Nanopore Sequencing for Investigation of the Human Epigenome

by

Isac Lee

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

May, 2020

Abstract

The invention of next generation sequencing (NGS) revolutionized the field of human genetics, providing a practical way to study genetics in a genome-wide fashion. With the ability to sequence the whole genome came the ability to observe the epigenome, the patterns and interactions of epigenetic signatures across the whole genome. Over the years, existing epigenetic methods have been adapted to NGS, and new approaches of using NGS to study the epigenome have been demonstrated. Now NGS-based studies are gold standards in studying the human epigenome.

Despite the advances in our understanding of the human epigenome, many gaps still remain. The epigenome is composed of numerous chemical components, and the observed phenotype is a result of complex interactions of these epigenetic features. The epigenome is also highly heterogeneous, further complicating the analysis of epigenetic interactions. In addition, limitations of our understanding of the genome, e.g. repetitive elements, centromeric regions, and large-scale genomic rearrangements, is translated directly to the lack of understanding of their epigenome.

The emergence of nanopore sequencing has opened new doors for studying the human genome and epigenome. Here I describe how I adapted NGS

epigenomic methods to nanopore sequencing, and how this furthers our knowledge of the epigenome. I first present the application of nanopore sequencing in observing multiple layers of epigenetic information and uncovering epigenetic patterns across the genome. Then I show the utility of the long read in studying the epigenome, observing allele-specific epigenetic patterns on single-molecules. Lastly, I implemented targeted sequencing on nanopore to generate deep sequencing coverage in select genomic regions and its usage in variant detection.

Primary Reader : Winston Timp, Ph.D.

Secondary Reader : Nickolas Papadopoulos, Ph.D.

Thesis Committee : Nickolas Papadopoulos, Ph.D., James Taylor, Ph.D., Kasper Daniel Hansen Ph.D., Winston Timp, Ph.D.

Preface

"Trust no one."

This is one of many pop culture references of Winston, my advisor. He says this to his mentees in context of our research, telling us to be critical of our own work as well as other's. This is only one of many things that shows the level of rigor that Winston expects in our science. Over the years of my graduate work, his mentorship has shaped me to strive for perfection in my work. I am thankful for the mentorship of Winston, being patient through my learning process, being approachable yet transparent in exchanging feedback, and sincerely caring for my well being both professionally and personally.

I am also grateful for the mentorship of my thesis committee. Nick Papadopoulos, the chair of my committee, refocused me to the biological motivation of my work. Kasper Hansen spent hours of his time to give me advice on my research methods as well as professional growth. James Taylor enrouraged me to be proud of my work and reach out to him about any concerns.

My friends and family were essential in helping me endure through the PhD years. Having great relationships with lab colleagues made going to lab fun, and the burger Wednesdays at PJ's and happy hours at The Dizz were important refreshers from lab work. My friends allowed me to connect with

the world in the midst of the busy-ness. My parents, sister, and brother-in-law emotionally supported me even though I am far away from home.

Last but not least, I thank my fiancée (wife as of June 2020) Sandra for her support in numerous ways. She is always on my side and supports everything I do, but she is also not hesistant to criticize me. She encourages me to do things in advance, and thanks to this I was able to find an internship as well as my post-graduate job. I can't imagine how I would have survived the last few years without her support.

Thank you everyone for being important parts of my PhD life.

Table of Contents

Abstract ii

Preface iv

Table of Contents vi

List of Tables xi

List of Figures xii

1 Introduction 1

2 Methylation and Accessibility Profile Analysis Using nanoNOMe 13

2.1 Abstract 13

2.2 Introduction 14

2.3 Results 17

2.3.1 Development of nanopore methylation calling model . 17

2.3.1.1 Methylation training and testing samples generation 17

2.3.1.2 Training the CpG + GpC dual methylation caller 18
2.3.1.3 Testing the methylation caller 20
2.3.2 nanoNOMe : Nanopore sequencing of Nucleosome Occupancy and Methylome 24
2.3.3 Comparison of nanoNOMe with conventional methodologies 28
2.3.4 Global epigenomic analysis of gene promoters and repetitive elements 33
2.3.5 Visualization and differential region detection 36
2.3.6 Comparative epigenomic analysis of breast cancer model 38
2.4 Discussion 44
2.5 Methods 44
2.5.1 Methylation training and testing set generation 44
2.5.2 Validation of DNA methylation by bisulfite sequencing 45
2.5.3 Processing of bisulfite sequencing data 45
2.5.4 nanopolish methylation training for dual CpG/GpC methylation calling 46
2.5.5 Cell Culture 47
2.5.6 Nucleosome footprinting via GpC methyltransferase . 47
2.5.7 Nanopore sequencing 48
2.5.8 Data Processing (basecalling, alignment, and structural variant calling) 49

2.5.9 Nanopolish methylation calling for dual CpG/GpC methylation . . . 49
2.5.10 Comparison of nanoNOMe with conventional methodologies . . . 50
2.5.11 Metaplot Analysis . . . 51
2.5.12 Enrichment analysis of differential epigenetic regions on genomic contexts . . . 52
2.6 Acknowledgments . . . 52
2.7 Supplementary Material . . . 53

3 Allele-specific and single-molecule epigenomic analysis 63
3.1 Abstract . . . 63
3.2 Introduction . . . 64
3.3 Results . . . 66
3.3.1 Co-occurrence of accessibility patterns to observe cis-regulatory interactions . . . 66
3.3.2 Improving single-molecule accessibility measurements using a smoothing estimator . . . 67
3.3.3 Resolving regulatory protein binding on individual reads 69
3.3.4 Single-molecule combinatorial promoter epigenetic states 73
3.3.5 Protein binding in association with promoter epigenetic state . . . 76

3.3.6 Allele-specific methylation and chromatin accessibility in X chromosome inactivation . . . 79
3.3.7 Genome-wide allele-specific epigenome analysis 80
3.3.8 Allele-specific epigenomics in heterozygous structural variations . . . 84
3.4 Discussion . . . 86
3.5 Methods . . . 87
3.5.1 Calculating and piling up co-occurrence of accessibility and inaccessibility . . . 87
3.5.2 Estimating single-molecule accessibility calls using a smoothing estimator . . . 88
3.5.3 Predicting regulatory protein binding from closed runs 88
3.5.4 Predicting combinatorial promoter epigenetic states on individual reads . . . 89
3.5.5 Interactions of promoter epigenetic states and protein binding . . . 90
3.5.6 Haplotype Assignment and Allele-Specific Methylation Analysis . . . 90
3.5.7 Bresat cancer cell line analysis . . . 91
3.6 Acknowledgments . . . 92
3.7 Supplementary Material . . . 92

4 Targeted sequencing on nanopore sequencing platform 103

4.1 Abstract . . . 103
4.2 Introduction . . . 104
4.3 Results . . . 105
4.3.1 Solution-phase hybridization capture nanopore sequencing . . . 105
4.3.2 Detecting nucleotide variations using targeted sequencing 108
4.4 Discussion . . . 110
4.5 Methods . . . 113
4.5.1 Agilent Sureselect XT Targeted Enrichment . . . 113
4.5.2 Data preprocessing . . . 114
4.5.3 Variant detection using targeted nanopore data . . . 114
4.6 Acknowledgments . . . 114
4.7 Supplementary Material . . . 115

5 Discussion and Conclusion **119**

Curriculum Vitae **124**

List of Tables

2.1 Sequencing metrics of nanoNOMe sequencing data 28

2.2 Nanopore sequencing yields of testing and training samples . 53

2.3 Relative accuracy and call rates for notable genomic contexts . 54

2.4 Individual nanopore sequencing run metrics of nanoNOMe samples 56

2.5 Summary of structural variations detected in breast cell lines . 57

4.1 Nanopore and Illumina targeted sequencing metrics 107

4.2 SNV detection metrics in the targeted regions of interest for three approaches 109

List of Figures

1.1 Overview of dissertation work 7

2.1 Overview of methylation calling using nanopolish 18

2.2 Overview of methylation control sample generation 19

2.3 Fractions of methylated sites in training and testing samples . 20

2.4 Current modulations in training samples 21

2.5 Receiver Operating Characteristic of methylation calling . . . 23

2.6 Performance of CpG and GpC dual methylation calling 25

2.7 Overview of nanoNOMe Assay 27

2.8 GC-content bias of coverage in sequencing methods 29

2.9 Distributions of genome-wide sequencing coverages 30

2.10 Fraction of low-complexity regions that were robustly mapped 31

2.11 Comparison of nanoNOMe profiles with WGBS, ATAC-seq, and DNAse-seq 32

2.12 Metaplots analysis of CpG methylation and nucleosome positioning on CTCF binding sites 33

2.13 NanoNOMe frequency analysis at transcription start sites . . . 35

2.14 NanoNOMe frequency analysis in repetitive elements 36
2.15 Detecting differentially methylated regions 38
2.16 Detecting differentially accessible regions 39
2.17 Bulk genome-wide differential methylation and accessibility analysis on breast cancer models 41
2.18 Enrichment of differential epigenetic regions in various genomic contexts 42
2.19 Epigenetic differences on an insertion only present in MCF-7 and MDA-MB-231 43
2.20 Distributions of current modulation in select 6-mers 54
2.21 Sequence context dependence of methylation calling accuracy 55
2.22 Per-Region nanoNOMe frequency in repetitive elements . . . 56
2.23 Structural variations and differential epigenetics 58

3.1 Matrix of accessibility co-occurrence on individual reads . . . 68
3.2 GpC accessibility kernel estimation on single reads 70
3.3 Single-molecule accessibility at CTCF binding regions 71
3.4 Per-read plot of methylation and accessibility on a CTCF binding site 72
3.5 Comparison of protein-binding predictions with ChIP-seq signals 73
3.6 Metaplots of CTCF binding sites stratified by ChIP-seq peak and read-level protein binding 74
3.7 Single-read epigenetic assessment on transcription start sites . 75

3.8 Clustering reads based on promoter combinatorial epigenetic state . . . 76
3.9 Comparisons of predicted TF-binding with respect to promoter epigenetic states . . . 78
3.10 Per-read plot of methylation and accessibility on a gene promoter with protein binding . . . 79
3.11 Read-level comparative epigenomic analysis of breast cancer model . . . 79
3.12 Haplotype phasing results on GM12878 nanoNOMe data . . . 81
3.13 Allele-specific epigenetics in X chromosome inactivation . . . 82
3.14 Enrichment of allele-specific differential epigenetic regions . . . 83
3.15 Allele-specific per-read methylation and accessibility of an ZNF597 . . . 84
3.16 CpG methylation in heterozygous structural variations . . . 85
3.17 Single-molecule closed run lengths at CTCF binding sites . . . 92
3.18 Assessment of read-level combinatorial epigenetic states of TSS 93
3.19 Results of Haystack Bio motif enrichment on candidate protein binding sites . . . 94
3.20 Differentially methylated and differentially accessible regions between alleles in GM12878 . . . 95
3.21 Genome-context enrichment of allele-specific differential epigenetic regions . . . 96

3.22 Allele-specific epigenetic comparison of heterozygous structural variations 97

4.1 Overview of solution-phase hybridization capture 106

4.2 Nanopore and Illumina sequencing coverage of the capture regions 107

4.3 SNV analysis using targeted nanopore sequencing 109

4.4 Structural Variation detection in targeted nanopore sequencing of Pancreatic Ductal Adenocarcinoma 111

4.5 Structural Variation detection in targeted nanopore sequencing of NA12878 115

Chapter 1

Introduction

When Conrad Waddington first coined the term "epigenetics" in 1942, he used it to broadly describe the mechanisms and processes linking the genotype and the phenotype (Waddington, 1942). As the field matured, this definition of epigenetics has evolved as well, now commonly being defined as potentially heritable features of the DNA other than the actual sequence that affect the function of the genome (Murrell, Rakyan, and Beck, 2005; Bernstein, Meissner, and Lander, 2007). In turn, the epigenome refers to the patterns and interactions of epigenetic features across the whole genome (Beck, Olek, and Walter, 1999). Despite the recent advances in the field and our better understanding of the epigenome, we still lack a clear understanding of how the many facets of the epigenome complement and interact with each other to regulate gene expression(Crews and Gore, 2014; Ng and Bird, 1999; Bell and Felsenfeld, 2000).

DNA methylation is a covalent modification of a DNA nucleotide residue by addition of a methyl group to the base. In mammals, the most prominent type of DNA methylation is cytosine methylation in CG dinucleotide contexts,

simply referred to as CpG methylation (Jones and Takai, 2001). DNA is also organized inside the nucleus by protein molecules called histones, and the DNA wrapped around complexes of histone molecules form packaging of chromatin termed nucleosomes. The organization of nucleosomes and the resulting differences in accessibility of chromatin govern the interactions of regulatory genomic elements, such as enhancers, promoters, and transcription factor binding motifs, with themselves and DNA-binding proteins (Klemm, Shipony, and Greenleaf, 2019). Factors that influence the organization of chromatin include, but are not limited to, covalent modifications on the N-terminal tails of histone molecules, interaction of DNA with the nuclear periphery, and binding of specific chromatin-remodeling proteins such as CTCF and cohesin (Kouzarides, 2007; Mattout-Drubezki and Gruenbaum, 2003; Rowley and Corces, 2018).

Because the epigenome influences which part of the genome is active, it has implications in many diseases and medical conditions (Robertson, 2005). One notable example of this is cancer: alterations in the epigenome are routinely observed in cancer (Hanahan and Weinberg, 2011). Global hypomethylation and hypermethylation of tumor supperssor gene promoters are some of the distinguishing epigenomic characteristics of cancer (Feinberg and Vogelstein, 1983; Baylin et al., 1986). Subsequent studies found that increase in variability and dysregulation of the epigenome contribute to these changes in the epigenome that promote tumorigenesis (Hansen et al., 2011; Landan et al., 2012; Timp and Feinberg, 2013). Another example is imprinting disorders:

because imprinting, or parent-of-origin specific expression of a gene, is maintained by the epigenome, conditions related to dysregulation of imprinting often have aberrant epigenetic signatures (Walter and Paulsen, 2003).

Discoveries of the role of the epigenome in diseases could not have been achieved without developments and advancements in epigenetic assays. Treatment of genomic DNA with methylation-sensitive restriction enzymes that have different cleaving capabilities depending on the methylation state of the recognition sequence, such as HpaII and HhaI, was one of the first methods used to detect CpG methylation, and is still a popular method to verify methylation states due to the robustness and low cost of the assay (Bird and Southern, 1978). In the early days of DNA sequencing, bisulfite conversion, in which methylated cytosine residues are protected from the conversion of unmethylated cytosine residues to uracil residues, became a popular method to couple DNA sequencing with methylation detection (Frommer et al., 1992). Then came microarray technology which, when coupled to methylation-sensitive restriction digestion or bisulfite conversion, allowed the first genome-wide assessment of DNA methylation using a single assay, in the sense that thousands of CpG sites across the genome could be examined (Gitan et al., 2002; Weber et al., 2005; Irizarry et al., 2008). Nucleosome organization and chromatin accessibility could be measured by treating intact nuclei with nucleases such as DNAse I and micrococcal nuclease Wu, 1980; Noll and Kornberg, 1977. The occupancy of DNA by proteins, i.e. histones, protects the DNA from these nucleases and only the open chromatin are cleaved, in the case of DNAse I, or digested away, in the case of micrococcal nuclease. The remaining short

strands of DNA could then be assessed using PCR or microarrays (Schones et al., 2008; Lee et al., 2007).

But, the entire landscape of genomics was revolutionized with the introduction of massively parallel DNA sequencing. Massively parallel sequencing, more commonly referred nowadays as Next Generation Sequencing (NGS), uses a fluidic device to simultaneously sequence millions of DNA strands at the same time (Margulies et al., 2005). Briefly, millions of DNA strands attach to the bottom of the flow cell and are amplified to generate clusters of DNA with strands within a cluster having identical sequences. Then the complementary strand of the DNA is synthesized one nucleotide at a time using reversible terminators, which are nucleotide molecules that prevent further extension. The nucleotide molecules also are labeled with fluorophores that emits light at a wavelength as the nucleotide gets incorporated. The multiple strands in each cluster amplifies the emitted signal, and the signals from millions of clusters are picked up by highly sensitive optical instruments. This process of *sequencing by synthesis* is repeated over and over until the full sequences of the strands are synthesized. This automated, highly parallel technology dramatically reduced the cost of DNA sequencing, making true whole genome sequencing of organisms with large genomes such as humans affordable. With the proliferation of NGS, many epigenetic assays were adapted to NGS to measure the epigenetic features across the genome, allowing true observation of the whole epigenome (Maunakea, Chepelev, and Zhao, 2010). Bisulfite sequencing, which had been applied on PCR and microarray technologies, was adapted to NGS as Whole Genome Bisulfite Sequencing (WGBS), allowing

genome-wide, single-base-resolution maps of human methylome (Lister et al., 2009). Existing chromatin assays were also adapted to NGS as whole genome enrichment sequencing methodologies, such as DNAse-seq and MNase-seq, to allow genome-wide assessment of chromatin organization (Boyle et al., 2008; Henikoff et al., 2011).

Though a tremendous advance, NGS still has limitations centered around the short lengths of its reads. This results in gaps in our knowledge; NGS-based epigenomic assays are unsuitable for studying repetitive elements and structural variations. Also, due to the highly heterogeneous nature of the epigenome, it has been difficult to directly link changes in the epigenome with phenotypic effects (Lai and Pugh, 2017; Buenrostro et al., 2015). Single-cell approaches have shown promise in navigating this heterogeneity, but because most human cells have two - or more, i.e. aneulploidy, - copies of each chromosome, these approaches still fail to resolve all of the heterogeneity. Lastly, the epigenome is composed of interactions of numerous components, but most epigenomic assays probe one of the epigenetic features (Bernstein, Meissner, and Lander, 2007).

Nanopore sequencing is a third generation of DNA sequencing technology that uses a transmembrane ion channel protein to characterize DNA molecules, where DNA sequences are detected in forms of electrolytic current modulation as molecules pass through a nanopore (Branton et al., 2009; Timp et al., 2010; Mikheyev and Tin, 2014). As the DNA strands pass through the protein molecule, which is large enough only for one single stranded DNA molecule at a time, the change in the current through the membrane is measured.

Multiple bases are within the central constriction of the pore at a given time ("k-mers") and the characteristic current signature given by each k-mer is deciphered (a.k.a base-called) into nucleotide sequences. Unlike sequencing by synthesis, nanopore sequencing measures the molecule directly and can detect covalent modifications on the DNA, such as CpG methylation (Simpson et al., 2017). Also, nanopore generates long DNA sequences (reads), which allows observing epigenetics of repetitive elements and structural variations (Jain et al., 2018; Norris et al., 2016). We can observe interactions of epigenetic features across the long length of single molecules, much like single-cell epigenomic assays, and further separate the molecules within a cell (phase) based on nearby non-homozygous variations (Gigante et al., 2019).

I leveraged nanopore sequencing to explore the human epigenome. I applied exogenous labeling to simultaneously measure CpG methylation and chromatin accessibility using nanopore sequencing, examining the epigenetic features in parts of the genome that have not been extensively studied. Then I exploited the long reads to explore the epigenome on an allele-specific and single-molecule resolution. Finally, I applied targeted nanopore sequencing, which, when coupled to the epigenomic methodologies, will allow large-scale epigenomic studies 1.1.

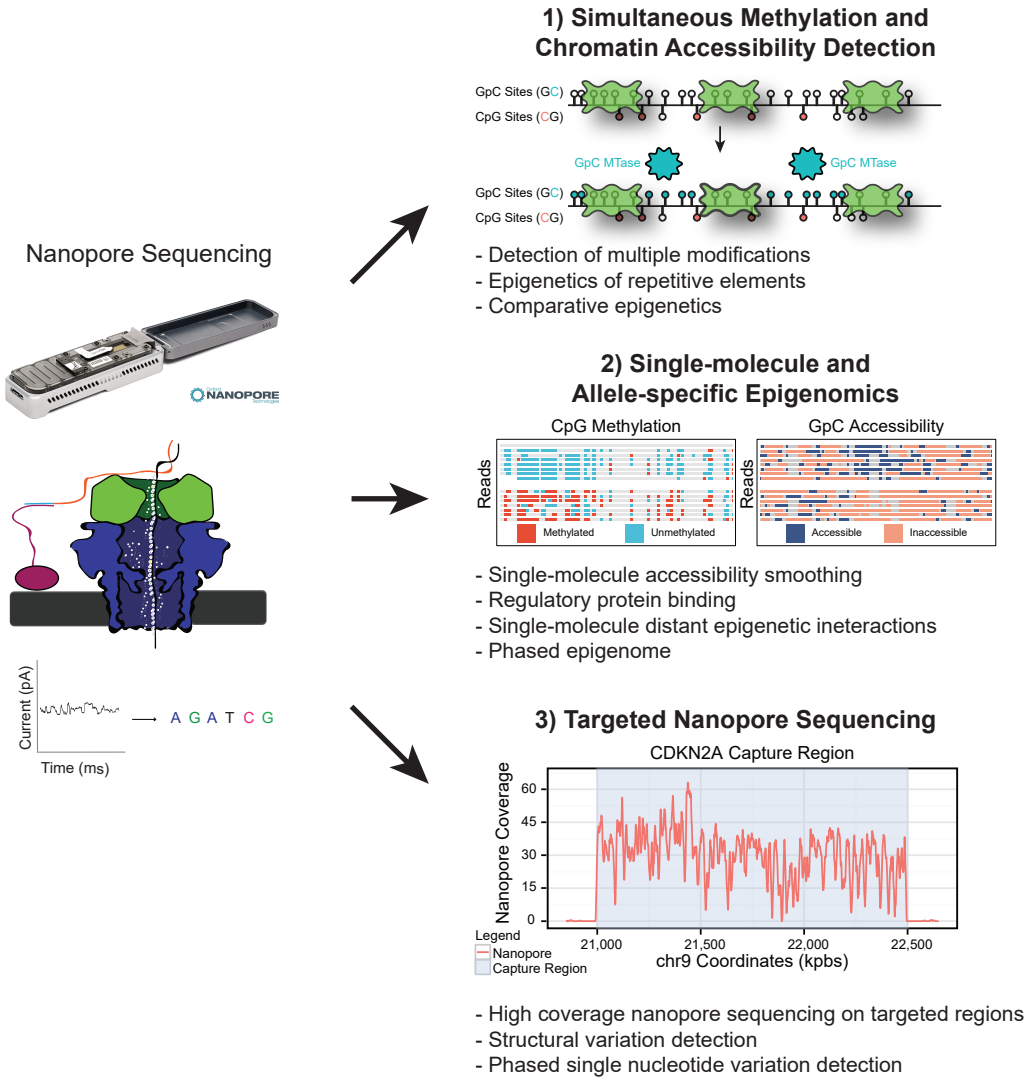


Figure 1.1: Overview of dissertation work.

References

Waddington, Conrad H (1942). "The epigenotype". In: *Endeavour* 1, pp. 18–20.

Murrell, Adele, Vardhman K Rakyan, and Stephan Beck (2005). "From genome to epigenome". en. In: *Hum. Mol. Genet.* 14 Spec No 1, R3–R10.

Bernstein, Bradley E, Alexander Meissner, and Eric S Lander (2007). "The mammalian epigenome". en. In: *Cell* 128.4, pp. 669–681.

Beck, S, A Olek, and J Walter (1999). "From genomics to epigenomics: a loftier view of life". en. In: *Nat. Biotechnol.* 17.12, p. 1144.

Crews, David and Andrea C Gore (2014). "Chapter 26 - Transgenerational Epigenetics: Current Controversies and Debates". In: *Transgenerational Epigenetics*. Ed. by Trygve Tollefsbol. Oxford: Academic Press, pp. 371–390.

Ng, H H and A Bird (1999). "DNA methylation and chromatin modification". en. In: *Curr. Opin. Genet. Dev.* 9.2, pp. 158–163.

Bell, A C and G Felsenfeld (2000). "Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene". en. In: *Nature* 405.6785, pp. 482–485.

Jones, P A and D Takai (2001). "The role of DNA methylation in mammalian epigenetics". en. In: *Science* 293.5532, pp. 1068–1070.

Klemm, Sandy L, Zohar Shipony, and William J Greenleaf (2019). "Chromatin accessibility and the regulatory epigenome". en. In: *Nat. Rev. Genet.* 20.4, pp. 207–220.

Kouzarides, Tony (2007). "Chromatin modifications and their function". en. In: *Cell* 128.4, pp. 693–705.

Mattout-Drubezki, A and Y Gruenbaum (2003). "Dynamic interactions of nuclear lamina proteins with chromatin and transcriptional machinery". en. In: *Cell. Mol. Life Sci.* 60.10, pp. 2053–2063.

Rowley, M Jordan and Victor G Corces (2018). "Organizational principles of 3D genome architecture". en. In: *Nat. Rev. Genet.* 19.12, pp. 789–800.

Robertson, Keith D (2005). "DNA methylation and human disease". en. In: *Nat. Rev. Genet.* 6.8, pp. 597–610.

Hanahan, Douglas and Robert A Weinberg (2011). "Hallmarks of cancer: the next generation". en. In: *Cell* 144.5, pp. 646–674.

Feinberg, A P and B Vogelstein (1983). "Hypomethylation distinguishes genes of some human cancers from their normal counterparts". en. In: *Nature* 301.5895, pp. 89–92.

Baylin, S B, J W Höppener, A de Bustros, P H Steenbergh, C J Lips, and B D Nelkin (1986). "DNA methylation patterns of the calcitonin gene in human lung cancers and lymphomas". en. In: *Cancer Res.* 46.6, pp. 2917–2922.

Hansen, Kasper Daniel, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A Irizarry, and Andrew P Feinberg (2011). "Increased methylation variation in epigenetic domains across cancer types". en. In: *Nat. Genet.* 43.8, pp. 768–775.

Landan, Gilad, Netta Mendelson Cohen, Zohar Mukamel, Amir Bar, Alina Molchadsky, Ran Brosh, Shirley Horn-Saban, Daniela Amann Zalcenstein, Naomi Goldfinger, Adi Zundelevich, Einav Nili Gal-Yam, Varda Rotter, and Amos Tanay (2012). "Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues". en. In: *Nat. Genet.* 44.11, pp. 1207–1214.

Timp, Winston and Andrew P Feinberg (2013). "Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host". en. In: *Nat. Rev. Cancer* 13.7, pp. 497–510.

Walter, Jörn and Martina Paulsen (2003). "Imprinting and disease". en. In: *Semin. Cell Dev. Biol.* 14.1, pp. 101–110.

Bird, A P and E M Southern (1978). "Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from Xenopus laevis". en. In: *J. Mol. Biol.* 118.1, pp. 27–47.

Frommer, M, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, and C L Paul (1992). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 89.5, pp. 1827–1831.

Gitan, Raad S, Huidong Shi, Chuan-Mu Chen, Pearlly S Yan, and Tim Hui-Ming Huang (2002). "Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis". en. In: *Genome Res.* 12.1, pp. 158–164.

Weber, Michael, Jonathan J Davies, David Wittig, Edward J Oakeley, Michael Haase, Wan L Lam, and Dirk Schübeler (2005). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation

in normal and transformed human cells". en. In: *Nat. Genet.* 37.8, pp. 853–862.

Irizarry, Rafael A, Christine Ladd-Acosta, Benilton Carvalho, Hao Wu, Sheri A Brandenburg, Jeffrey A Jeddeloh, Bo Wen, and Andrew P Feinberg (2008). "Comprehensive high-throughput arrays for relative methylation (CHARM)". en. In: *Genome Res.* 18.5, pp. 780–790.

Wu, C (1980). "The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I". en. In: *Nature* 286.5776, pp. 854–860.

Noll, M and R D Kornberg (1977). "Action of micrococcal nuclease on chromatin and the location of histone H1". en. In: *J. Mol. Biol.* 109.3, pp. 393–404.

Schones, Dustin E, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao (2008). "Dynamic regulation of nucleosome positioning in the human genome". en. In: *Cell* 132.5, pp. 887–898.

Lee, William, Desiree Tillo, Nicolas Bray, Randall H Morse, Ronald W Davis, Timothy R Hughes, and Corey Nislow (2007). "A high-resolution atlas of nucleosome occupancy in yeast". en. In: *Nat. Genet.* 39.10, pp. 1235–1244.

Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors". en. In: *Nature* 437.7057, pp. 376–380.

Maunakea, Alika K, Iouri Chepelev, and Keji Zhao (2010). "Epigenome mapping in normal and disease States". en. In: *Circ. Res.* 107.3, pp. 327–339.

Lister, Ryan, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor

Ruotti, A Harvey Millar, James A Thomson, Bing Ren, and Joseph R Ecker (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences". en. In: *Nature* 462.7271, pp. 315–322.

Boyle, Alan P, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford (2008). "High-resolution mapping and characterization of open chromatin across the genome". en. In: *Cell* 132.2, pp. 311–322.

Henikoff, Jorja G, Jason A Belsky, Kristina Krassovsky, David M MacAlpine, and Steven Henikoff (2011). "Epigenome characterization at single base-pair resolution". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.45, pp. 18318–18323.

Lai, William K M and B Franklin Pugh (2017). "Understanding nucleosome dynamics and their links to gene expression and DNA replication". en. In: *Nat. Rev. Mol. Cell Biol.* 18.9, pp. 548–562.

Buenrostro, Jason D, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf (2015). "Single-cell chromatin accessibility reveals principles of regulatory variation". en. In: *Nature* 523.7561, pp. 486–490.

Branton, Daniel, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, Stevan B Jovanovich, Predrag S Krstic, Stuart Lindsay, Xinsheng Sean Ling, Carlos H Mastrangelo, Amit Meller, John S Oliver, Yuriy V Pershin, J Michael Ramsey, Robert Riehn, Gautam V Soni, Vincent Tabard Cossa, Meni Wanunu, Matthew Wiggin, and Jeffery A Schloss (2009). "The potential and challenges of nanopore sequencing". In: *Nanoscience and Technology*. Co-Published with Macmillan Publishers Ltd, UK, pp. 261–268.

Timp, Winston, Utkur M Mirsaidov, Deqiang Wang, Jeff Comer, Aleksei Aksimentiev, and Gregory Timp (2010). "Nanopore Sequencing: Electrical Measurements of the Code of Life". en. In: *IEEE Trans. Nanotechnol.* 9.3, pp. 281–294.

Mikheyev, Alexander S and Mandy M Y Tin (2014). "A first look at the Oxford Nanopore MinION sequencer". en. In: *Mol. Ecol. Resour.* 14.6, pp. 1097–1102.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). "Detecting DNA cytosine methylation using nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 407–410.

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". en. In: *Nat. Biotechnol.* 36.4, pp. 338–345.

Norris, Alexis L, Rachael E Workman, Yunfan Fan, James R Eshleman, and Winston Timp (2016). "Nanopore sequencing detects structural variants in cancer". en. In: *Cancer Biol. Ther.* 17.3, pp. 246–253.

Gigante, Scott, Quentin Gouil, Alexis Lucattini, Andrew Keniry, Tamara Beck, Matthew Tinning, Lavinia Gordon, Chris Woodruff, Terence P Speed, Marnie E Blewitt, and Matthew E Ritchie (2019). "Using long-read sequencing to detect imprinted DNA methylation". en. In: *Nucleic Acids Res.* 47.8, e46.

Chapter 2

Methylation and Accessibility Profile Analysis Using nanoNOMe

Isac Lee, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J. Sedlazeck, Kasper D. Hansen, Jared T. Simpson, Winston Timp. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. bioRxiv 504993 (2018) doi:10.1101/504993

2.1 Abstract

Probing epigenetic features on long molecules of DNA has tremendous potential to advance our understanding of the epigenome. In this section, we evaluate CpG methylation and chromatin accessibility simultaneously on long strands of DNA using GpC methyltransferase to exogenously label open chromatin, coupled with nanopore sequencing technology. We describe the procedures we used to train and test the model for calling CpG and GpC

methylation simultaneously. We then describe nanopore sequencing of Nucleosome Occupancy and Methylome (nanoNOMe) on four human cell lines (GM12878, MCF-10A, MCF-7, MDA-MB-231), and demonstrate the ability to directly measure methylation and chromatin accessibility in genomic features such as structural variations and repetitive elements.

2.2 Introduction

With the proliferation of DNA sequencing technologies, methods have been developed for examining nuclear organization, protein binding site occupancy, chromatin accessibility, and methylation state using next generation sequencing (NGS). Cytosine methylation at CG dinucleotide contexts (CpG methylation) has been studied widely using bisulfite treatment coupled to DNA sequencing (Frommer et al., 1992). In bisulfite sequencing, native genomic DNA undergoes bisulfite conversion, wherein unmethylated cytosine residues are converted to uracil residues while methylated cytosine residues are protected from this conversion. The resulting DNA sequences are computationally parsed to determine the methylation state, and despite the reduction of DNA complexity due to the bisulfite conversion, advances in computational algorithms have made the process highly accurate and reproducible (Krueger and Andrews, 2011; Hansen, Langmead, and Irizarry, 2012).

Many methods for detecting chromatin accessibility and nucleosome positioning rely on the openness of accessible and nucleosome depleted chromatin to enzymatic treatments. DNAse-seq is a NGS adaptation of DNAse I

hypersensitivity assay, which uses DNAse I to selectively cleave nucleosome-depleted DNA and retrieve DNAse I hypersensitive sites (Boyle et al., 2008). By amplifying the cleaved, short strands of DNA, the resulting DNA library becomes enriched of the accessible chromatin, which is then sequenced to observe the enrichment. Similarly, ATAC-seq exploits the accessibility of cleaving enzymes to open chromatin; instead of DNAse I, ATAC-seq uses Tn5 transposases to insert a predefined sequence to the open chromatin after cleaving the site (Buenrostro et al., 2015). Then these predefined transposon sequences are used as the template sequences for PCR, thereby selectively amplifying the cleaved sites. The approach of MNase-seq is the opposite : instead of cleaving and enriching the open chromatin, MNase digests away the open chromatin, leaving only the inaccessible, nucleosome-occupied DNA to be amplified and sequenced (Henikoff et al., 2011).

NOMe-seq uses a methyltransferase enzyme to exogenously label accessible chromatin (Kelly et al., 2012). The methyltransferase enzyme used in this methodology, M. CviPI GpC methyltransferase, methylates cytosine residues at GpC sites (GC dinucleotide contexts). While cytosine methylation is endogenously present in mammalian genomes, because the vast majority of endogenous cytosine methylation specifically occurs in CpG sites (CG dinucleotide contexts), the GpC methylation can be separated as the exogenous labeling motif. After bisulfite conversion and sequencing, the resulting data contains both the endogenous CpG methylation and exogenously labeled GpC accessibility. Therefore, NOMe-seq permits simultaneous evaluation of

endogenous CpG methylation and nucleosome occupancy. Another distinguishing feature of NOMe-seq with regard to measuring chromatin accessibility is that it does not rely on enrichment by PCR amplification. The usage of covalent DNA modification rather than PCR enrichment makes this easily adaptable to nanopore sequencing.

Nanopore sequencing is suitable for epigenomic studies because of its ability to detect covalent modifications, as these modifications change the chemical structure of the DNA molecule and hence the modulation of the current by the molecule. We and others have previously shown that endogenous CpG methylation can be accurately called with nanopore data (Simpson et al., 2017; Rand et al., 2017). Another advantage of nanopore sequencing is that it generates long reads, which allows a deeper analysis into long-range patterns on individual molecules. More recently, this technology was applied to exogenous labeling of chromatin accessibility, similar to in NOMe-seq, in S. Cerevisiae, a unicellular eukaryotic model organism without endogenous methylation (Shipony et al., 2020; Wang et al., 2019). To further the application of nanopore sequencing in studying the epigenome, here we present nanopore sequencing of Nucleosome Occupancy and Methylome (nanoNOMe), where we label mammalian cells which have endogenous CpG methylation with exogenous GpC modifications at accessible sites. We are able to take advantage of the long read lengths (>10kb) generated by nanopore sequencing to read the CpG methylation and chromatin accessibility across stretches of genomic regions at the single molecule level.

2.3 Results

2.3.1 Development of nanopore methylation calling model

2.3.1.1 Methylation training and testing samples generation

The current gold standard method for detecting methylation on nanopore sequencing data is nanopolish (**Figure 2.1**) (Simpson et al., 2017). Nanopolish employs a hidden Markov model (HMM) to detect cytosine methylation based on electrical current signatures (events) corresponding to groups of nucleotide sequences (k-mers). The HMM uses a table of event level distributions, termed a pore model, characteristic to every 6-mer to predict the methylation state of the 6-mer. The original methylation pore model was designed to call cytosine methylation at CG dinucleotide contexts (CpG methylation), which is the endogenous methylation present in mammalian DNA (Simpson et al., 2017). Extending the methylation caller to additionally call cytosine methylation at GC dinucleotide context (GpC methylation) required building a new model able to handle four possible states for each 6-mer: 1) methylated at CpG contexts, 2) methylated at GpC contexts, 3) methylated at both CpG and GpC contexts, and 4) unmethylated at both contexts. Building a new model required new training samples representing each methylation state. We generated the training samples using combinations of M.SssI (CpG methyltransferase) and M. CviPI (GpC methyltransferase) on unmethylated (PCR amplified) Escherichia coli (E. coli) genomic DNA (gDNA) (**Figure 2.2**). We also generated testing samples in the same fashion on NA12878 human lymphoblast cell line gDNA. Testing samples are generated separately, on gDNA

of an organism with more complex genome to ensure during testing that the model has not been overtrained. To verify that the methylation was successful, the training and testing sets were subjected to bisulfite sequencing on Illumina Sequencing. After separating the cytosine motifs by the dinucleotide context and calculating the fraction of methylated sites, we observed near-complete methylation specifically at desired motifs (**Figure** 2.3).

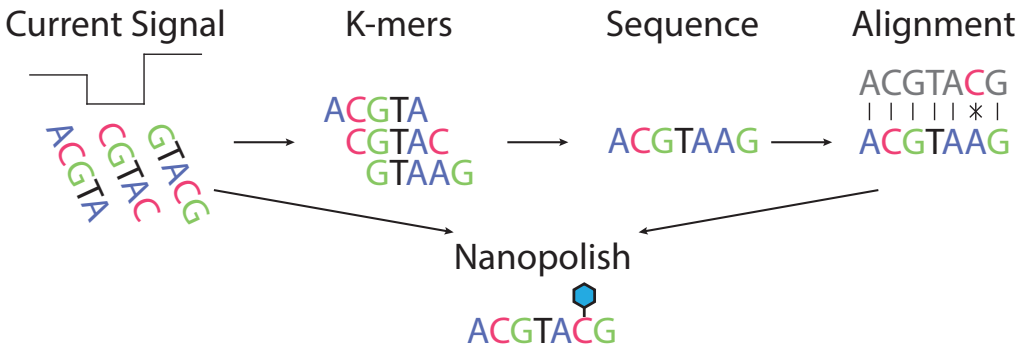


Figure 2.1: Overview of methylation calling using nanopolish. The current signal is converted in to the sequence, and then mapped to a reference genome. The resulting alignment information and the raw current signal are both used to determine the methylation state of the data

2.3.1.2 Training the CpG + GpC dual methylation caller

Once the quality of methylation samples were verified via Illumina sequencing, we sequenced the samples on a minION flowcell of nanopore sequencing. We generated an average of 7.8 Gb of DNA sequences per sample, amounting to an average of 1,300X of E. coli genome coverage and 3X of human genome coverage (**Supplementary Table** 2.2). After processing the training data and aligning to E. coli reference sequence, we tabulated the current modulations (mA) for each alignment to a methylation motif (CpG and/or GpC) in the reference sequence. We observed that the different positions of the methylation

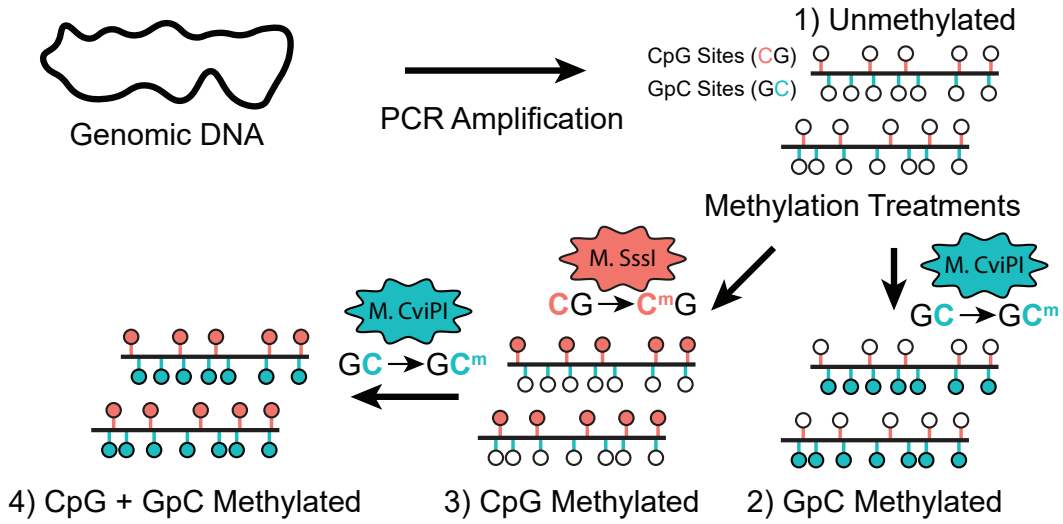


Figure 2.2: Overview of methylation control sample generation. Genomic DNA is amplified by PCR, generating (1) fully unmethylated DNA. On aliquots of the unmethylated DNA, combinations of CpG and GpC methylation treatments are performed using M. SssI for CpG and M. CviPI for GpC methylation, generating (2) GpC methylated, (3) CpG methylated, and (4) CpG and GpC methylated samples

within 6-mers modulate the current to different extents, which we can use to discriminate between CpG and GpC methylation in addition to methylated and unmethylated states (**Figure** 2.4**a**, more examples in **Supplementary Figure** 2.20). We then tested for the dependence of the current modulation on the position of the methylation within the 6-mer and found that the current deviation from unmethylated 6-mers is weak when the methylation motif is at the edge (1st and 6th position) and the strongest when it is on the 5th position of the 6-mer (**Figure** 2.4**b**).

We then used the E. coli nanopore sequencing data to train the dual CpG/GpC methylatin pore model. For each 6-mer in the reference sequence, the posterior probability of an observed event was learned using the forward/backward algorithm on a simplified HMM. For each 6-mer, we fit a

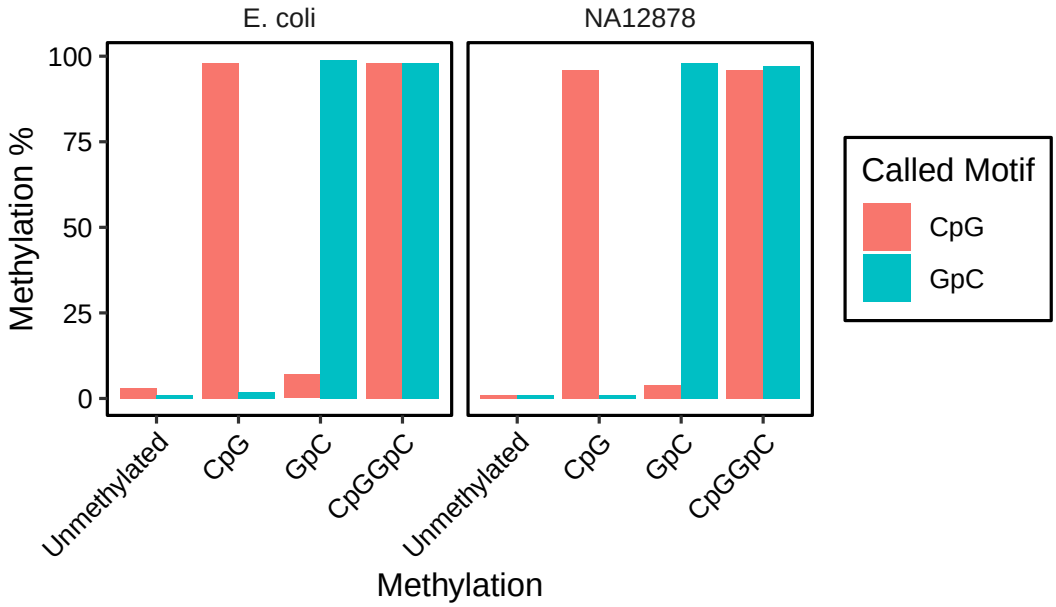


Figure 2.3: Fractions of methylated sites in training and testing samples. Fractions of CpG and GpC sites in methylation training (E. coli) and testing (NA12878) sets that were called as methylated by bisulfite sequencing.

Gaussian model to the distribution of the current modulations using the expectation-maximization algorithm, with the contribution of each observation weighted by the calculated posterior probability. Fitting the Gaussian model for each 6-mer in each dataset, we generated the pore models for the four states of methylation.

2.3.1.3 Testing the methylation caller

We then used the methylation pore model to call methylation on the testing sample dataset. To be able to call the two methylation motifs simultaneously and also discriminate the two motifs, the methylation calling module in nanopolish was modified. The first step of methylation calling is grouping nearby CpG and GpC sites that are less than 5 bp apart in order to minimize

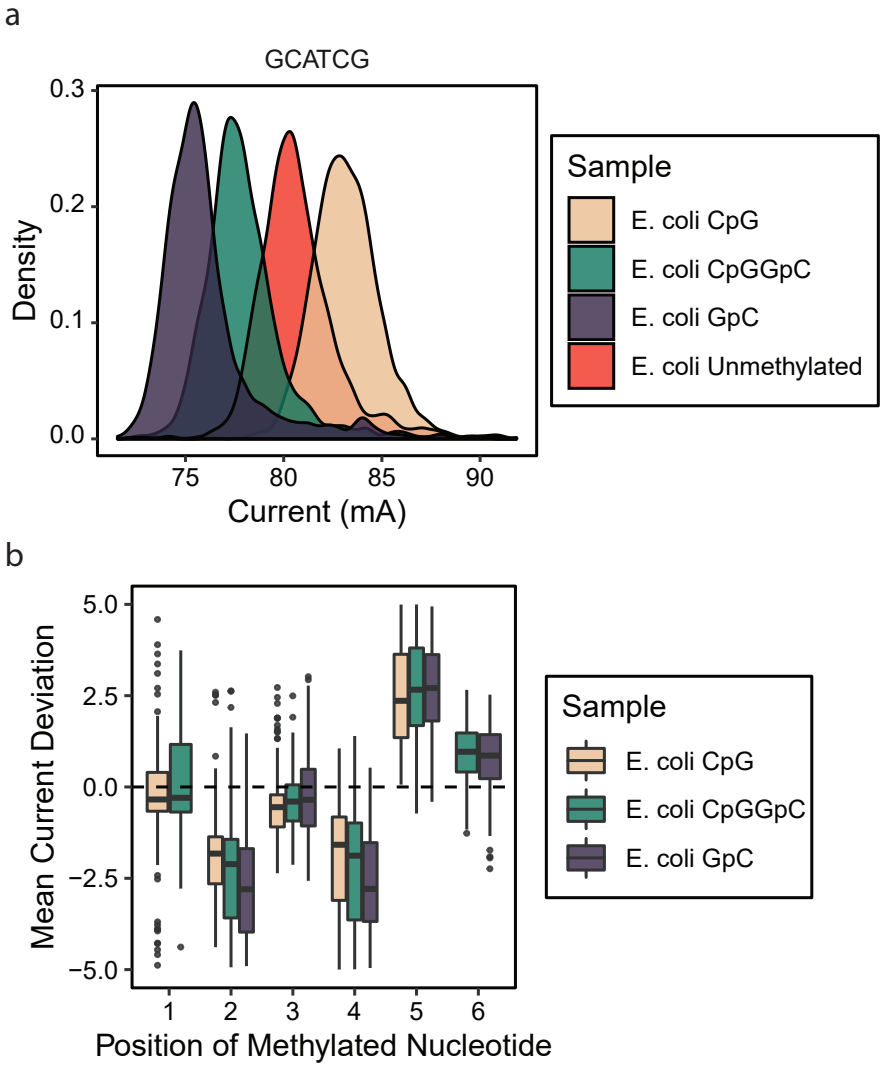


Figure 2.4: Current modulations in training samples. The ability of nanopore sequencing to distinguish cytosine methylation at CpG and GpC contexts is shown by **(a)** an example of shift in current modulations in the four training samples at GCATCG 6-mer and **(b)** examining current level shifts depending on the placement of the methylation on a 6-mer

the influence of adjacent methylation motif in the current modulation of another motif. We then calculate a likelihood for combinations of the grouped sites being methylated or unmethylated (either no sites methylated, all CpGs

methylated, all GpCs methylated, or all sites both contexts methylated), using the k-mer states trained in the previous section, with a profile HMM. Finally, separate ratios of log-likelihoods (LLR) are calculated for the two methylation motifs :

$$LogLikRatio(X_{CpG}) = log(\mathcal{L}(X_{CpG}) + \mathcal{L}(X_{CpGGpC})) - log(\mathcal{L}(X_{GpC}) + \mathcal{L}(X_{Unmethylated}))$$

$$LogLikRatio(X_{GpC}) = log(\mathcal{L}(X_{GpC}) + \mathcal{L}(X_{CpGGpC})) - log(\mathcal{L}(X_{CpG}) + \mathcal{L}(X_{Unmethylated}))$$

To benchmark the dual methylation detection, we called methylation on the NA12878 testing sample data. After nanopore sequencing the same samples, we first confirmed that full methylation does not decrease mappability of the reads in nanopore sequencing (**Supplementary Table** 2.2). We observed that the fraction of high-quality reads did not decrease in the methylated samples in comparison to the unmethylated sample. We then tested the performance of nanopolish by calulating the Receiver Operating Characteristic (ROC) curves : we applied a range of LLR thresholds to bin the continuous LLR into the binary state of methylated/unmethylated calls and compared the predictions to the true state for each singularly methylated and the unmethylated data **Figure** 2.5). Both CpG and GpC ROC had high areas under the curve (AUC) (0.91 for CpG and 0.98 for GpC), validating the performance of the methylation

calling model.

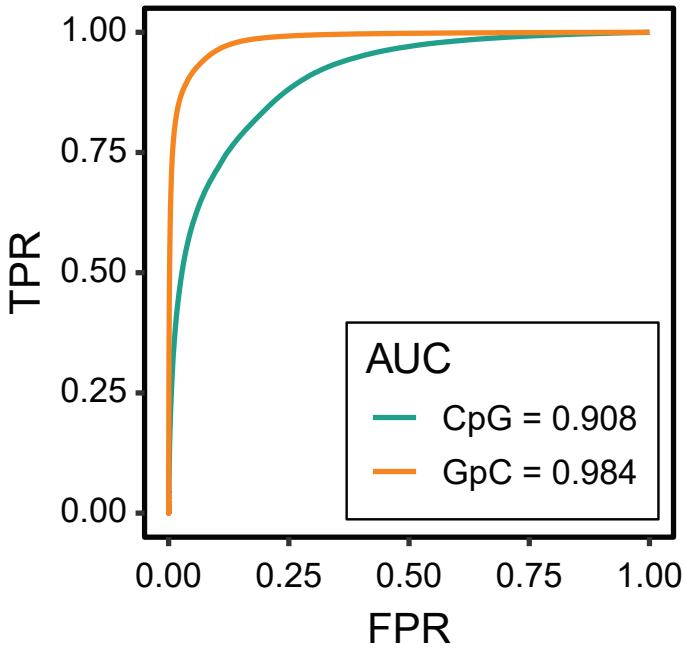


Figure 2.5: Receiver Operating Characteristic of methylation calling. ROC curve of methylation calling for a range of LLR thresholds on NA12878 modified with CpG and GpC methylation.

To choose the LLR thresholds for determining the binary state of methylation (methylated vs unmethylated) from the continuous LLRs, we separated the LLRs of both the E. coli and GM12878 methylation controls based on the true states, into groups of methylated calls and unmethylated calls. From the distributions of these LLRs, we calculated the LLR that would allow 5% false calls : i.e. top 5th percentile of LLRs for unmethylated calls and bottom 5th percentile of LLRs for methylated. Then, to make the thresholds symmetric between methylated and unmethylated calls, we averaged the absolute value of the two thresholds and applied a ceiling function to the nearest half. As a result, we chose a threshold of 1.5 for calling CpG methylation (LLR < -1.5 is

unmethylated, and > 1.5 is methylated, and values between are uncalled) and a threshold of 1 for GpC methylation. Using these cutoff values on NA12878 testing data, 91% of CpG calls correctly identified methylation at the 72% of CpGs that pass the threshold and 96% of GpCs calls correctly identified at the 93% that pass the threshold (**Figure 2.6a**). This is a conservative estimate of our accuracy because these metrics were calculated with the assumption that the methylated input was 100% methylated, whereas the bisulfite sequencing data indicated incomplete (96-98%) enzymatic methylation in this testing set. We then looked at CpG and GpC methylation in all of the testing samples, and verified that the presence of methylation at a different motif does not affect the accuracy of the calls for a given methylation motif (**Figure 2.6b**). Motif analysis confirmed that the k-mers that had a higher ratio of ambiguously called to correctly called did not have a sequence bias beyond the GCG motif, which is already excluded from our analysis (**Supplementary Figure 2.21**). Genome context analysis confirmed that neither the fraction of sites called nor the fraction of accurate calls was dependent on the genomic context (**Supplementary Table 2.3**).

2.3.2 nanoNOMe : Nanopore sequencing of Nucleosome Occupancy and Methylome

We adapted the NOMe-seq protocol (Kelly et al., 2012) to exogenously label open chromatin with GpC methylation and apply it to nanopore sequencing, terming this modified method nanopore sequencing of Nucleosome Occupancy and Methylome (nanoNOMe) (**Figure 2.7**). The methylation treatment of intact nuclei results in GpC methylation only at unoccupied, open regions

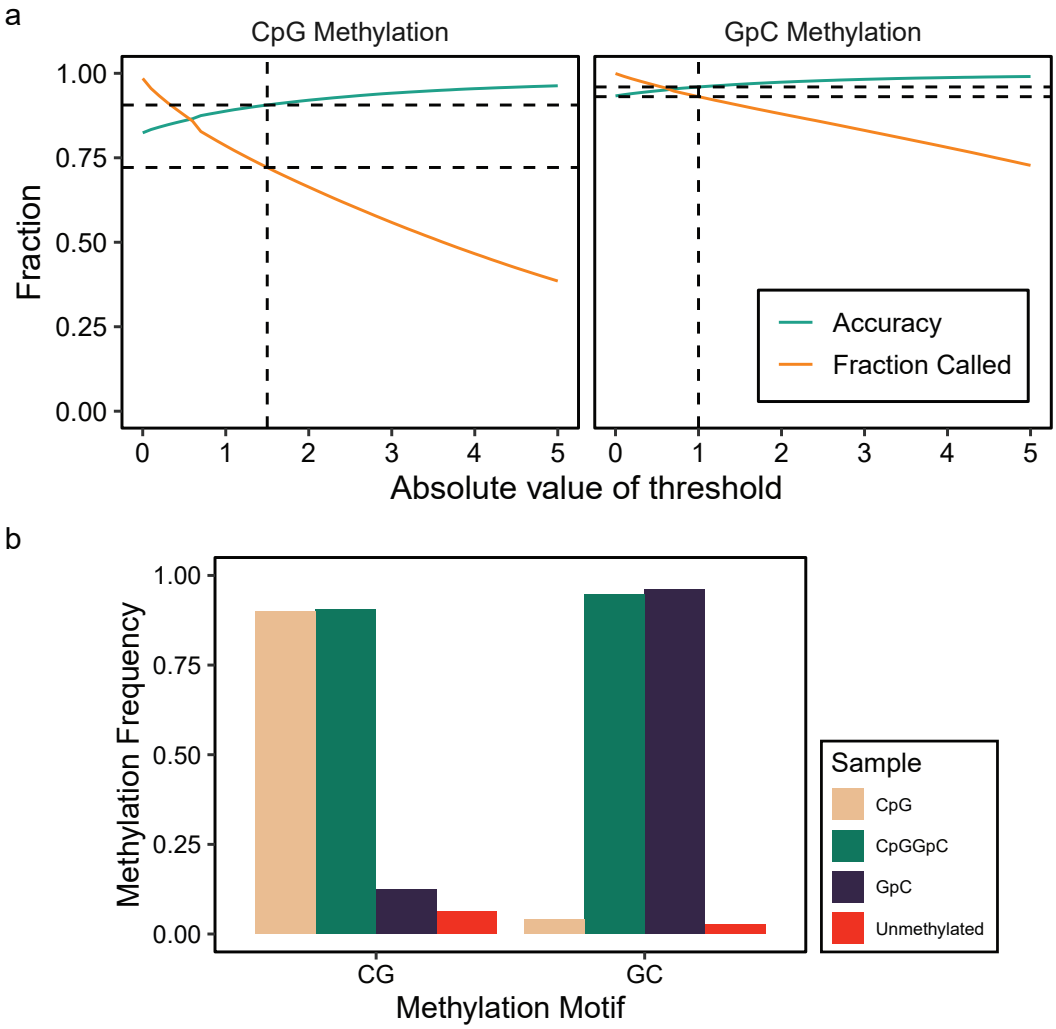


Figure 2.6: Performance of CpG and GpC dual methylation calling. (a) Fraction of k-mers passing the threshold filter and the fraction of k-mers from which methylation was correctly called for a range of thresholds. **(b)** methylation frequencies for calling both CpG and GpC methylation in the NA12878 testing samples.

of the genome. Briefly, intact nuclei were extracted from cells by gentle lysis, followed by methylation with GpC methyltransferase. After purification of DNA from these nuclei by phenol:chloroform extraction and ethanol precipitation, instead of traditional bisulfite conversion, we proceeded to ligation-based library preparation for nanopore sequencing (ONT). Because

nanopore sequencing discriminates methylated cytosine residues directly, bisulfite conversion is unnecessary. However, to preserve the modifications, we cannot amplify the DNA which necessitates a higher (1-2µg) input amount of DNA. After sequencing, basecalling, and alignment, we applied our CpG + GpC dual methylation model to detect methylation in both CpG and GpC contexts. Methylation at cytosine residues in GCH contexts was used as a measure of chromatin accessibility and cytosine residues in HCG contexts as a measure of endogenous methylation (methylation measurements in GCG cytosine residues were excluded from analysis). In describing GpCs state, a methylated GpC was interpreted as an accessible mark, and unmethylated as inaccessible.

We performed nanoNOMe on a well-characterized GM12878 lymphoblast cell line (Zook et al., 2016). The advantage of performing nanoNOMe on GM12878 is threefold : 1) Because GM12878 is one of the most deeply characterized cell line in ENCODE, there are multiple high-quality datasets of epigenetic assays on GM12878 (ENCODE Project Consortium, 2012). We can use these datasets to benchmark nanoNOMe and test whether the measurements from nanoNOMe are comparable to orthogonal techniques. 2) In addition to assays that probe similar aspects o the epigenome as nanoNOMe, we can couple the nanoNOMe data with other regulatory datasets on GM12878, such as RNA-seq and ChIP-seq (ENCODE Project Consortium, 2012), and observe the association of DNA methylation and chromatin accessibility with other aspects of genome regulation. 3) Because parent-of-origin single nucleotide variations (SNVs) have been characterized in GM12878 (Eberle et al., 2016),

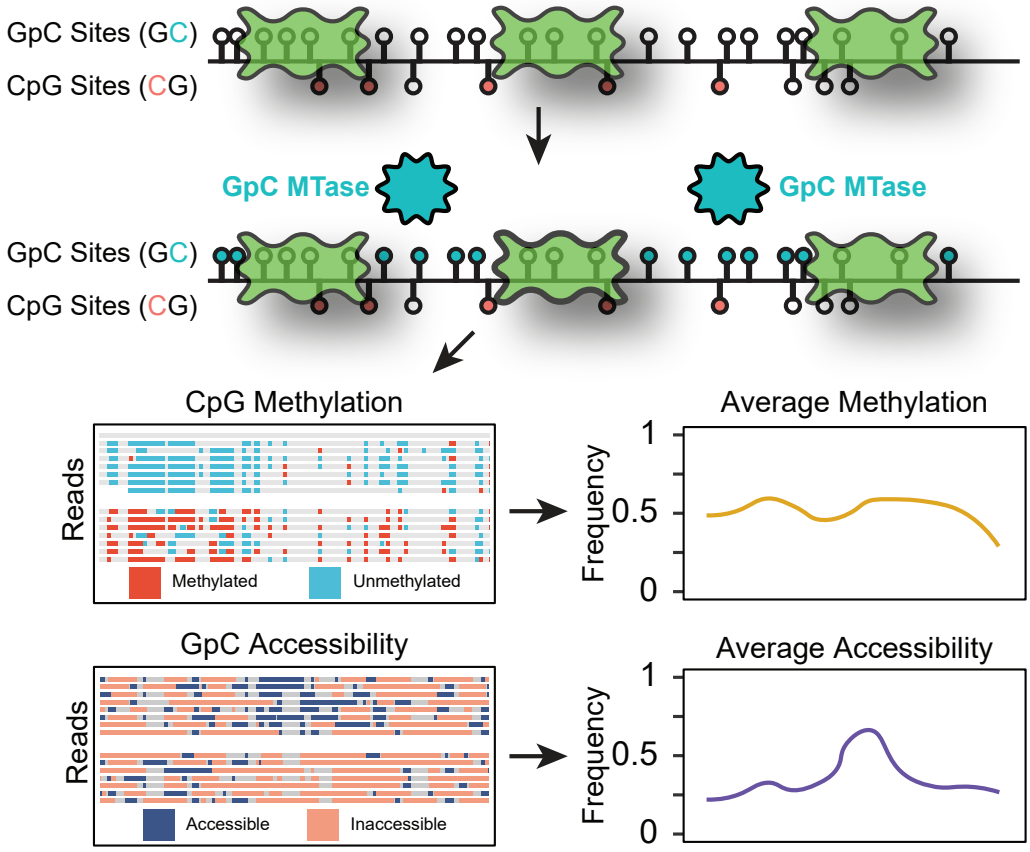


Figure 2.7: Overview of nanoNOMe Assay. After purifying intact nuclei, sample is methylated at GpC motifs to serve as the mark for accessibility. CpG methylation and GpC accessibility can be simultaneously measured via nanopore sequencing.

the reads that originate from parts of the genome with heterozygous SNVs can be divided into the parent of origin and the allele-specific epigenome can be studied. We generated 250 Gb (103X coverage) of mapped sequencing data from 15 flowcells (12 minION and 3 PromethION), with an N50 read length of 14,000 bp. (**Table 2.1, Supplementary Table 2.4**).

2.3.3 Comparison of nanoNOMe with conventional methodologies

To examine the mappability of nanoNOMe and whether the exogenous labeling causes any biases in mappability, we compared genomic coverage of the resulting GM12878 nanoNOMe data to WGBS from a previous study (100X coverage, ENCODE accession ENCSR890UQO) (ENCODE Project Consortium, 2012) and whole-genome nanopore sequencing of GM12878 (36X coverage, ENA accession code PRJEB23027) (Jain et al., 2018). One of the most notorious pitfalls of NGS is the bias of coverage based on the percentage of Guanine and Cytosine residues (GC content) in the reference sequence (Olova et al., 2018; Ji et al., 2014). We divided non-ambiguous portions of the human reference sequence into 200 bp bins and calculated the GC content of each bin along with the number of reads that were mapped to the bin in each of the datasets (**Figure 2.8**). Whereas WGBS coverage was biased by GC content of the reference sequence bin and increased with increasing GC content, nanoNOMe and nanopore WGS both showed consistent coverages that was not dependent on GC content. We also plotted distributions

Cell	Number of flowcells	Number of raw reads (M)	Total raw bases (Gb)	Aligned reads (M)	Aligned bases (Gb)	Average Coverage	N50 length
GM12878	12 + 3 Plon*	32.0	298.3	26.4	256.9	103	14,020
MCF-10A	9	9.4	81.6	7.7	72.4	27	11,501
MCF-7	11	9.0	76.8	7.5	69.1	26	13,025
MDA-MB-231	9	8.0	82.4	7.0	74.9	28	13,507

* PromethION flowcell (all other were MinION)

Table 2.1: Sequencing metrics of nanoNOMe sequencing data. NanoNOMe was performed on four cell lines using multiple runs of MinION, GridION, or PromethION sequencing and pooled to generate one data set per cell line.

of coverages irrespective of the GC content and compared to Poisson distribution estimates using means of the coverages (**Figure** 2.9). When reads are distributed randomly across the genome, the coverage follows a Poisson distribution (Lander and Waterman, 1988). We observed that while nanoNOMe and nanopore WGS coverages closely line up with their respective Poisson distribution estimates, the distribution of WGBS coverages were wider than its Poisson estimate (Standard Errors: WGBS=1.26, nanoNOME=0.16, Nanopore WGS=0.16) , indicating that the coverage is more variable in WGBS.

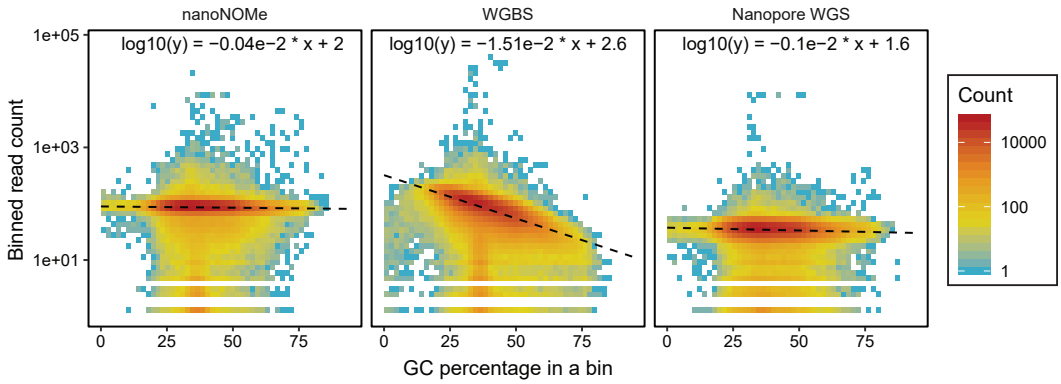


Figure 2.8: GC-content bias of coverage in sequencing methods. Binned coverage versus GC percentage in corresponding bins of nanoNOMe, WGBS (ENCODE), and WGS by nanopore sequencing, along with regression models representing the degree of dependence of coverage on GC-content in the form of the slope of the regression model.

We then examined regions that are poorly mappable via short reads. We focused on regions that had 10 or more reads with mapping quality (MAPQ) less than 5, based on bowtie2 alignments in the bismark pipeline, to determine regions that have low mappability in WGBS. These regions covered 132 Mb of the human genome, consisting of 57,982 distinct regions with an average size of 2.3 kb. The coverage of high quality reads (MAPQ > 20) for nanoNOMe

was between the 5th and 95th percentile of genome coverage (between 67X and 116X) for 44% of these regions with median coverage of 114X, in contrast to only 7% in WGBS (between 23X and 168X) with an abnormally high median coverage of 582X compared to the overall median coverage of 100X. This demonstrates that long read sequencing, and specifically nanoNOMe, does not suffer from mismapping of reads to poorly mappable regions and is able to cover these regions of poor mappability. As a second metric, we examined genomic contexts known to be difficult to map: CpG islands, repetitive elements (SINE, LINE, LTR), and satellite regions. For each genomic context, we identified regions that had coverages between the 5th and 95th percentile of genome coverage. We observed that nanoNOMe had higher fractions of such regions for all five genomic contexts, especially LINE and CGI, demonstrating that nanoNOMe is more robust in aligning to low complexity regions and interrogating epigenetics in these regions (**Figure** 2.10).

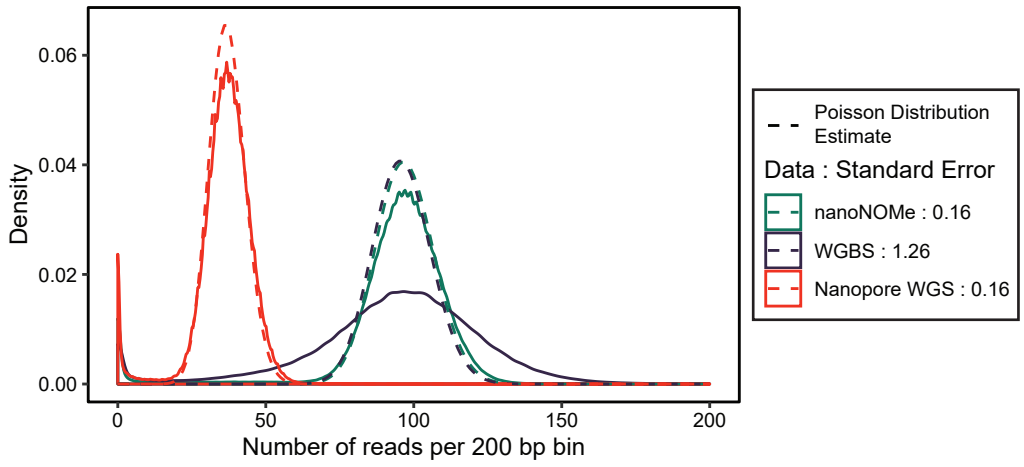


Figure 2.9: Distributions of genome-wide sequencing coverages. Density distributions of binned coverages show heavy GC bias and deviation from Poisson distribution in WGBS but not in either of the nanopore-based data.

We next assessed the performance of nanoNOMe in resolving endogenous cytosine methylation and chromatin accessibility. On each CpG site in the human reference sequence that had data in both WGBS and nanoNOMe, we calculated the frequency of methylation and performed a pairwise comparison of the frequencies from the two methods (**Figure 2.11a**). The two datasets had a high correlation across the genome (Pearson correlation of 0.92), validating that the endogenous CpG methylation signal from nanoNOMe are not negatively affected by the exogenous labeling. Using the GpC accessibility frequency, we detected peaks of accessibility across the genome in nanoNOMe (see Results 2.3.5). We detected a total of 69,305 peaks, and 85% (58,742) of these peaks overlapped with peaks called by ATAC-seq and/or DNAse-seq (**Figure 2.11b**), demonstrating that accessibility signal from nanoNOMe agrees well with ATAC-seq and DNAse-seq data.

One of the regulatory features that has a clear association with CpG methylation and nucleosome occupancy is binding sites of CTCF transcription factor

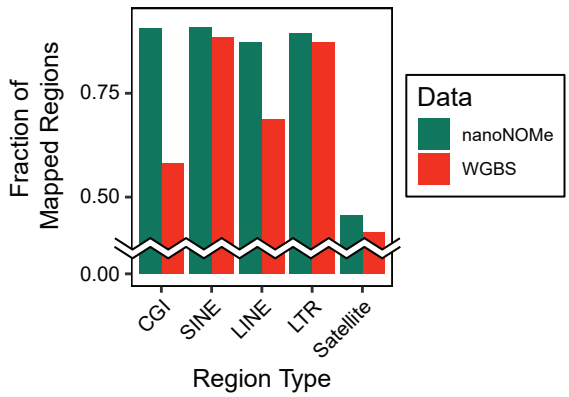


Figure 2.10: Fraction of low-complexity regions that were robustly mapped. Comparison of fraction of low sequence complexity regions between WGBS and nanoNOMe that had coverage between 5th and 95th percentile of genome coverage.

(Kelly et al., 2012; Bell and Felsenfeld, 2000). We performed metaplot analysis on CTCF binding sites to examine the ability of nanoNOMe signals to observe the patterns of CpG methylation and nucleosome positioning in specific genomic contexts. Metaplots are plots of aggregated CpG methylation and GpC accessibility frequencies across the genome with respect to the distance from the center of a genomic feature. Because the frequencies are averaged across the genome in the feature of interest, only the patterns that are strongly associated with the genomic feature will be highlighted in the resulting metaplot. We generated metaplots of nanoNOMe CpG methylation and GpC accessibility with respect to CTCF binding sites, and verified the profiles by plotting the same metaplots using WGBS and MNase-seq data (**Figure 2.12**, ENCODE accession ENCSR890UQO and ENCSR000CXP). NanoNOMe metaplots closely agreed with respective orthogonal metaplots : in CpG methylation, the region surrounding the CTCF binding sites are consistently demethylated and

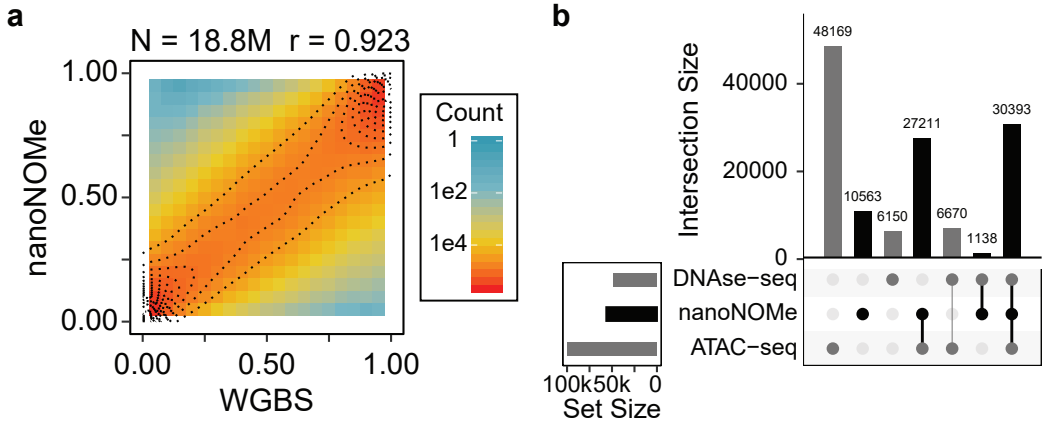


Figure 2.11: Comparison of nanoNOMe profiles with WGBS, ATAC-seq, and DNAse-seq. Validation of NanoNOMe profiles by **(a)** pair-wise comparison of per-CpG average methylation from nanoNOMe with WGBS in across the genome, and **(b)** intersections of accessibility peaks from nanoNOMe, DNAse-seq, and ATAC-seq.

accessible. Both epigenetic features show consistent oscillation of the signal with a width of 172 bp, indicating consistent positioning of mononucleosomes (Radman-Livaja and Rando, 2010). In addition, nanoNOMe captures the occupancy due to the CTCF protein binding, shown by the narrow decrease in accessibility at the center of the binding site.

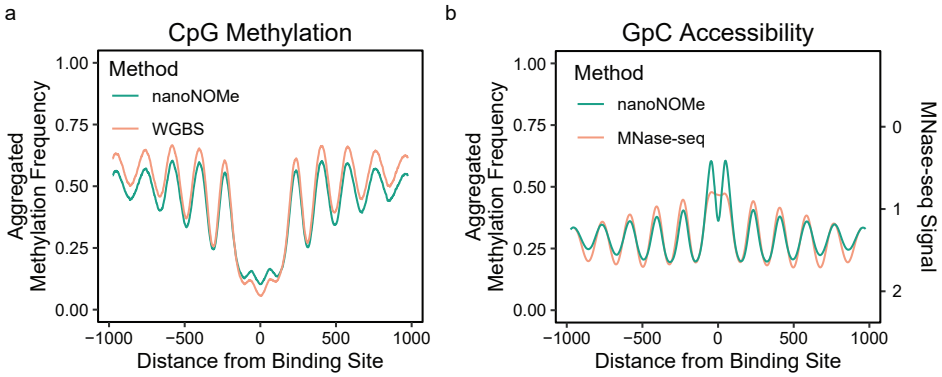


Figure 2.12: Metaplots analysis of CpG methylation and nucleosome positioning on CTCF binding sites. Aggregated frequency plots of **(a)** CpG methylation and **(b)** GpC accessibility as a function of distance to CTCF binding motifs in nanoNOMe, WGBS, and MNAse-seq.

2.3.4 Global epigenomic analysis of gene promoters and repetitive elements

To correlate nanoNOMe signals with chromatin states and gene activity, we performed metaplot analysis on transcription start sites (TSSs). We generated metaplots at TSSs with euchromatic (H3K4me3) and heterochromatic (H3K27me3) histone modifications using existing ChIP-seq data on GM12878 (**Figure** 2.13**a**, ENCODE accessions ENCSR057BWO and ENCSR000AKD). As expected, CpG methylation decreased and GpC accessibility increased at the TSS in promoters with active H3K4me3 marks, in contrast to the high CpG

methylation and low accessibility at promoters with repressive H3K27me3 marks. To observe the association of the two epigenetic features with gene activity, we separated the TSSs based on the expression quartiles of the corresponding gene transcripts, and observed an increase in chromatin accessibility and decrease in CpG methylation with increasing expression level (**Figure 2.13b**, ENCODE accessions ENCSR843RJV). We also examined the combination of the two features at TSSs, i.e. the combinatorial epigenetic states of gene promoters, in association with expression and found that the active genes are characterized by high accessibility and low methylation (concordantly active), and inactive genes are characterized by low accessibility and high methylation (concordantly inactive) promoter epigenetic states (**Figure 2.13c**).

We also characterized patterns of methylation and accessibility in repetitive elements in GM12878. We focused on LINE, LTR, Alu, and MIR, which are the four most abundantly annotated repetitive elements in the human reference genome. We compared distributions of per-CpG methylation in the repetitive regions to randomly shuffled regions of same width that do not overlap with the repetitive elements (**Figure 2.14a**). Interestingly, only Alu elements exhibited a difference in the global distribution of methylation, higher methylation than the rest of the genome, whereas the other elements did not show a deviation. This was consistent when comparing per-region average methylation instead of per-CpG methylation **Supplementary Figure 2.22**). Accessibility in repetitive elements were examined by comparing the number of accessibility peaks per megabase in these regions compared to the

number across the genome (**Figure 2.14b**). Peaks were depleted in all repetitive elements, especially in LINE and LTR regions, indicating that repetitive

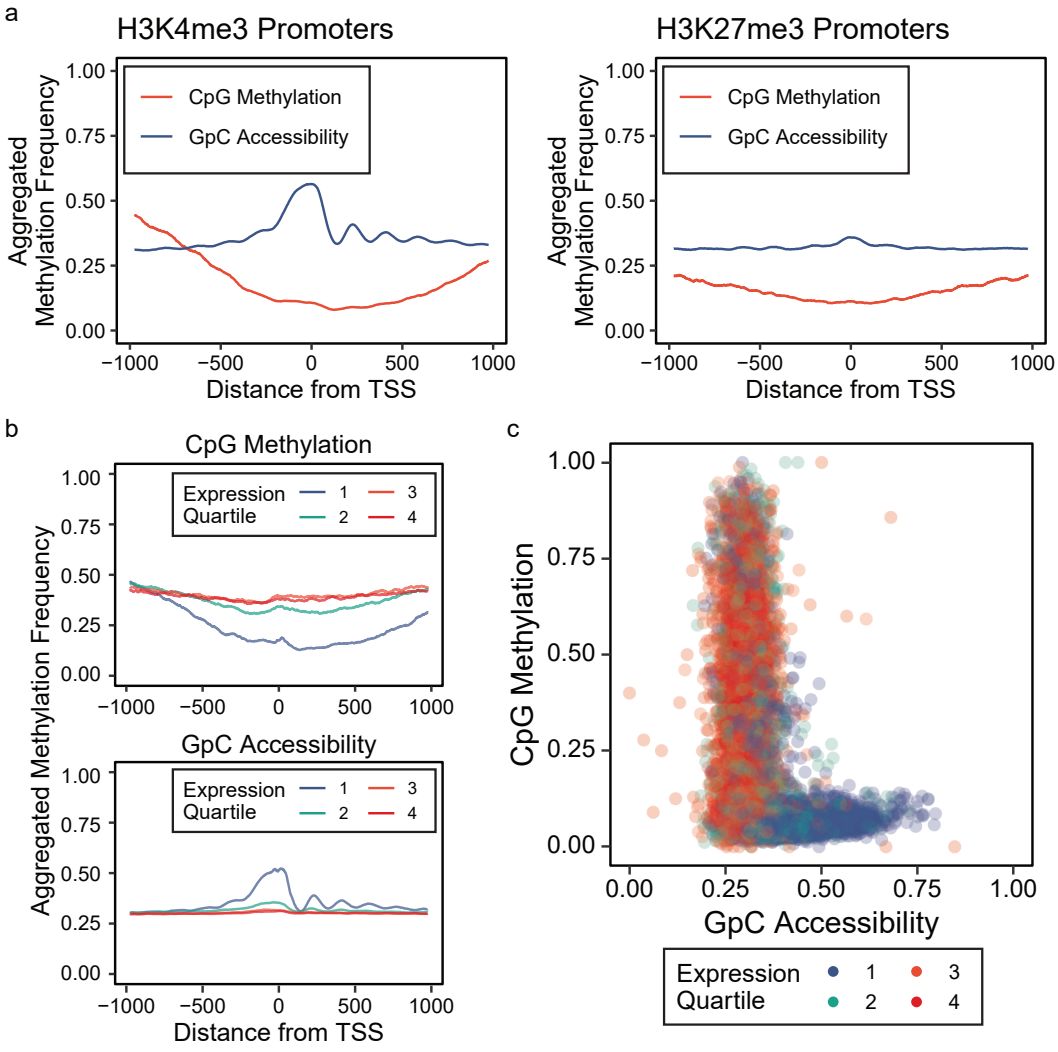


Figure 2.13: NanoNOMe frequency analysis at transcription start sites. (a) Metaplots of TSS with euchromatin (H3K4me3) and heterochromatin (H3K27me3) histone modifications within 1 kb of the TSS. **(b)** Metaplots of TSS divided up into expression quartiles of the corresponding gene transcripts (Descending order : 1st is highest and 4th is lowest expressing transcripts). **(c)** Pairwise scatter plot of average CpG methylation to GpC accessibility for 400 bp regions around each TSS, colored by the expression quartiles.

elements have decreased accessibility across the genome in GM12878.

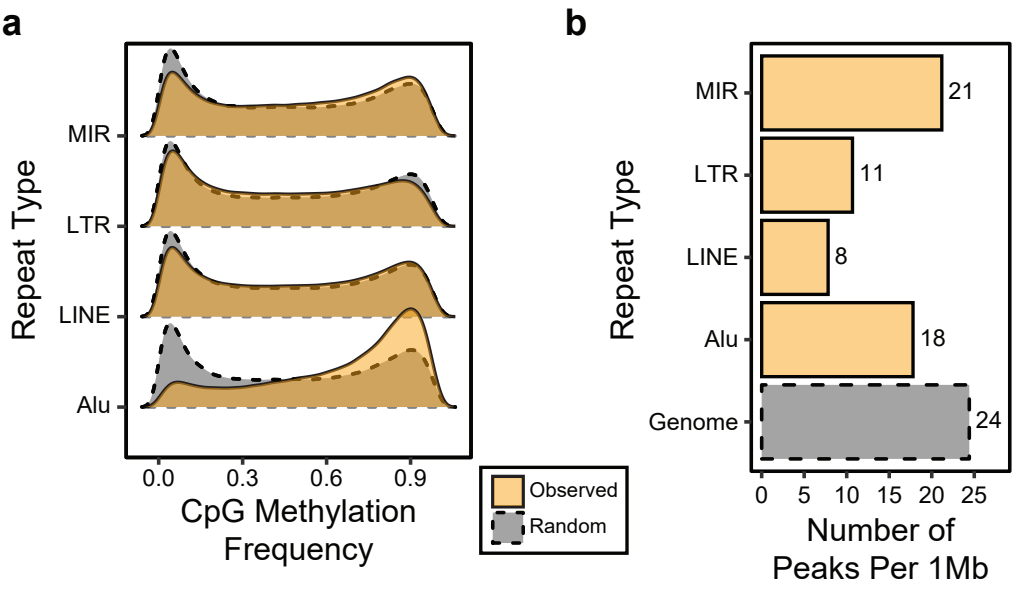


Figure 2.14: NanoNOMe frequency analysis in repetitive elements. (a) Distribution of observed per-CpG site methylation frequency in repetitive elements in comparison to random regions across the genome of the same lengths and **(b)** number of accessibility peaks per 1Mb of repetitive regions in comparison to the entire genome.

2.3.5 Visualization and differential region detection

For comparison and visualization of bulk methylation and accessibility, estimated profiles of measurements were calculated by fitting locally weighted generalized linear models across the genome. Previous studies have shown that DNA methylation is spatial well-correlated over distances less than 1 kb, and locally smoothing increases reproducibility and consistency of the methylation profile (Eckhardt et al., 2006; Hansen, Langmead, and Irizarry, 2012). For each CpG site, at least 50 nearby CpG sites and sites within 1 kb of the site were used to estimate the smoothed methylation frequency. GpC

accessibility profile was smoothed as well, with reduced windows of 100 bp and at least 10 nearby GpC sites to account for more rapid fluctuations in the signal in nucleosome free regions.

To find differentially methylated regions (DMRs) between two samples without replicates, the difference of the smoothed methylation frequencies between the two samples was calculated for each CpG site (**Figure 2.15**). Then, continuous regions with differences greater than 99th percentile of the per-site differences were selected as candidates for hypermethylation, and regions with differences less than the 1st percentile were selected as candidates for hypomethylation. Similarly, for DARs, we performed a one-sided Fisher's Exact test on raw counts of methylated and unmethylated calls on each candidate DMR. P-values were corrected using Benjamini-Hochberg correction, and regions with adjusted p-values less than 0.01 and widths greater than 100 bps were determined to be significant DMRs.

To find regions of high accessibility, continuous regions having smoothed accessibility greater than 99th percentile of the data were first selected (**Figure 2.16**). The significance of each accessible region was determined by performing a binomial test of the raw frequency of accessibility, with overall accessibility frequency as the null probability. The probabilities were corrected for multiple testing using Benjamini-Hochberg correction, and accessible regions with adjusted p-values less than 0.01 and widths greater than 50 bps were determined to be accessibility peaks. To determine differentially accessible regions (DARs) between two samples without replicates, we identified accessibility peaks that were present exclusively in one of the samples. For each candidate DAR,

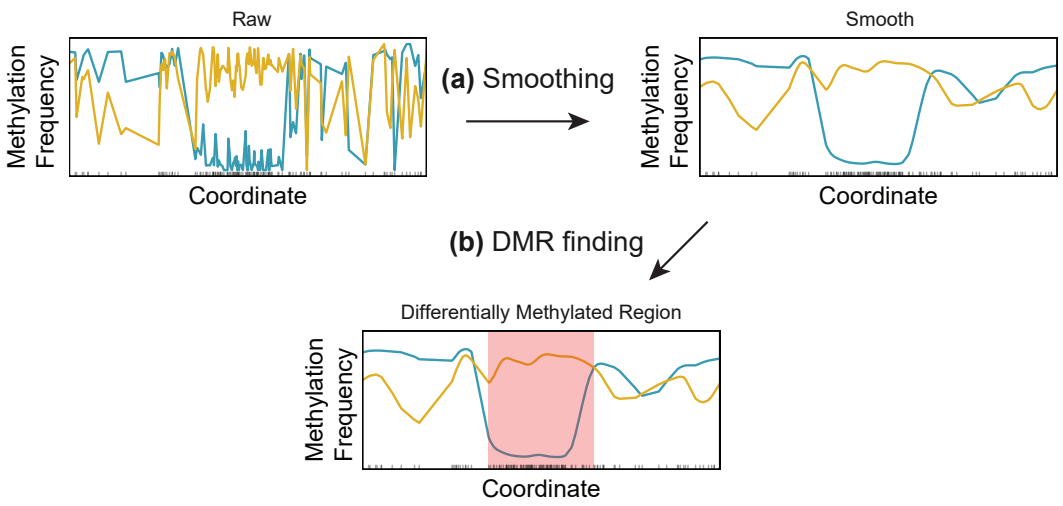


Figure 2.15: Detecting differentially methylated regions. **(a)** Methylation profiles are separately smoothed using locally weighted smoothing, then **(b)** the methylation is compared between two samples and continuous regions of significant differences are detected with Fisher's Exact test and Benjamini-Hochberg correction.

we performed a one-sided Fisher's Exact test on raw counts of accessible and inaccessible calls. P-values were corrected using Benjamini-Hochberg correction, and regions with adjusted p-values less than 0.01 were determined to be significant DARs.

2.3.6 Comparative epigenomic analysis of breast cancer model

We applied nanoNOMe to measure epigenetic differences between three well-characterized breast cell lines: MCF-7 (luminal breast carcinoma, ER+/PR+/HER2-) and MDA-MB-231 (basal breast carcinoma, ER-/PR-/HER2-) as two subtypes of breast cancer, and MCF-10A (fibrocystic disease) as the normal baseline (Holliday and Speirs, 2011; Messier et al., 2016). We performed the assay and generated >20X whole genome coverage of nanoNOMe data per cell

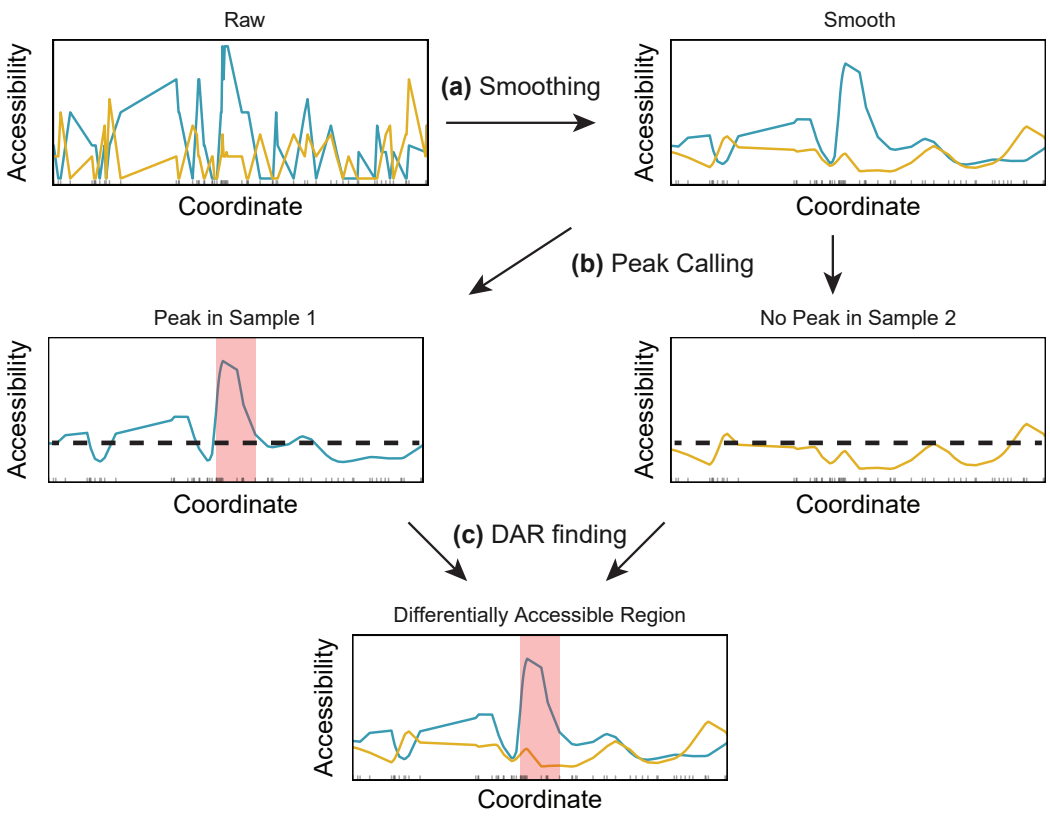


Figure 2.16: Detecting differentially accessible regions. (a) Accessibility profiles are separately smoothed using locally weighted smoothing, then **(b)** continuous regions of high accessibility are selected as peaks of accessibility. **(c)** Differentially accessible regions are detected by finding peaks that are exclusive to one of the two samples, and comparing the accessibility frequency in the regions between the two samples with Fisher's Exact test and Benjamini-Hochberg correction.

line (**Table 2.1, Supplementary Table 2.4**). We detected differentially methylated regions (DMRs) and differentially accessible regions (DARs) across the genome between MCF-10A and the two cancer subtypes (**Figure 2.17**). Both of the cancer subtypes had higher numbers of hypomethylated DMRs than hypermethylated DMRs (1.8-fold for MCF-7 and 7.6-fold for MDA-MB-231) suggesting global hypomethylation in the cancer subtypes, consistent with

previous findings that cancer cells are often globally hypomethylated in comparison to normal tissues (**Figure 2.17a**)(Ehrlich, 2009). In addition, though to a lesser extent than in methylation, more DARs pointed at higher accessibility in cancer cells than in MCF-10A (1.4-fold for MCF-7 and 1.3-fold for MDA-MB-231)(**Figure 2.17b**). Combinatorially, only a subset of DMRs and DARs coincided at the same genomic loci (8191 overlapping regions, 11% of DMRs and 6% of DARs), but coinciding DMRs and DARs were highly concordant, with hypermethylation coinciding with a decrease in accessibility and vice versa (Pearson correlation < -0.9), highlighting the complementary nature of differential methylation and accessibility (**Figure 2.17c**). We refer to these regions of concordance between a DMR and DAR as concordantly differential regions (CDR). Interestingly, more concordant differential regions indicated less activity (decreases of accessibility and increase in methylation) in the cancer subtypes, especially in MCF-7 (2.5-fold for MCF-7 and 1.2-fold for MDA-MB-231) (**Figure 2.17d**).

We then examined the genomic contexts of differential epigenetic regions by calculating the enrichment of DMRs, DARs, and CDRs in a number of genomic regions that are associated with regulatory functions (**Figure 2.18**). We found that CTCF binding regions, and transcription factor (TF) binding regions to a lesser extent, are highly enriched in all three types of differential regions in both MCF-7 and MDA-MB-231. In MCF-7, the DARs in these regulatory regions were more accessible and TF binding regions were more cncordantly active, indicating a global increase in affinity to regulatory proteins in MCF-7 (Rothenberg, 2014). In MDA-MB-231, more DARs in CTCF

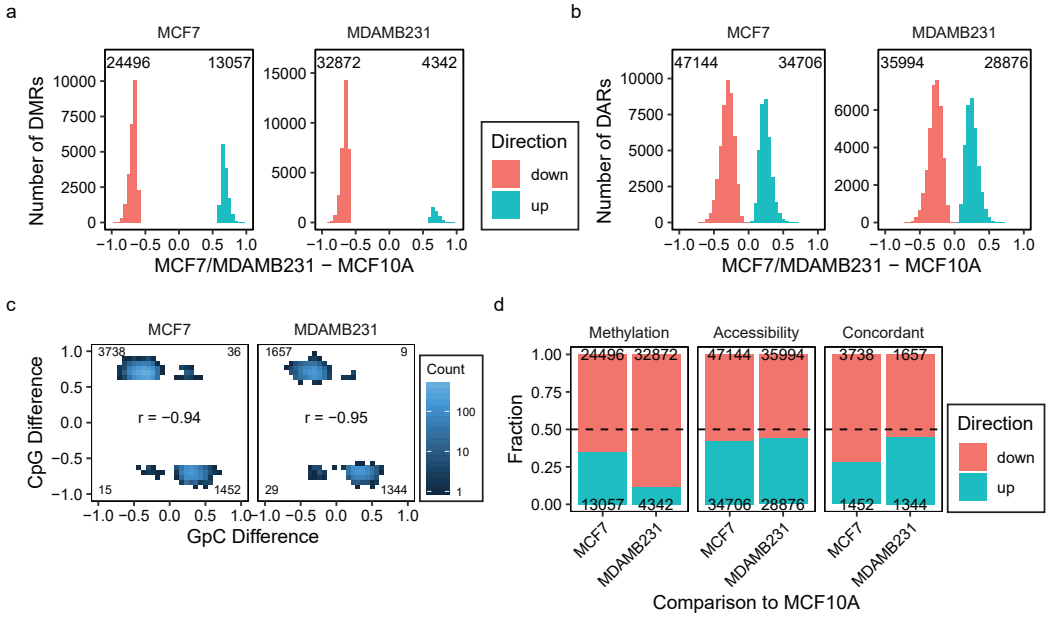


Figure 2.17: Bulk genome-wide differential methylation and accessibility analysis on breast cancer models. Histograms of the difference between MCF-7/MDA-MB-231 and MCF-10A **(a)** methylation in differentially methylated regions and **(b)** accessibility in differentially accessibility regions. **(c)** Comparison of average methylation to average accessibility in regions that have both differential methylation and accessibility, showing that in regions of significant methylation/accessibility difference, the two features are strongly epigenetically concordant. **(d)** Comparisons of the directions of DMRs, DARs, and concordantly differential regions. Dotted line is the 1:1 ratio.

binding regions were less accessible, more methylated, and less concordantly active, suggesting a global decrease in CTCF binding in MDA-MB-231 (Bell and Felsenfeld, 2000). None of the repetitive elements showed an enrichment of differential epigenetic regions, indicating that there is no difference in global activity of repetitive elements.

We also leveraged long reads to detect SVs and observe differences in epigenetic features near the SV breakpoints. Excluding SVs that occurred commonly between the three cell lines and those < 50 bp, we called a total of

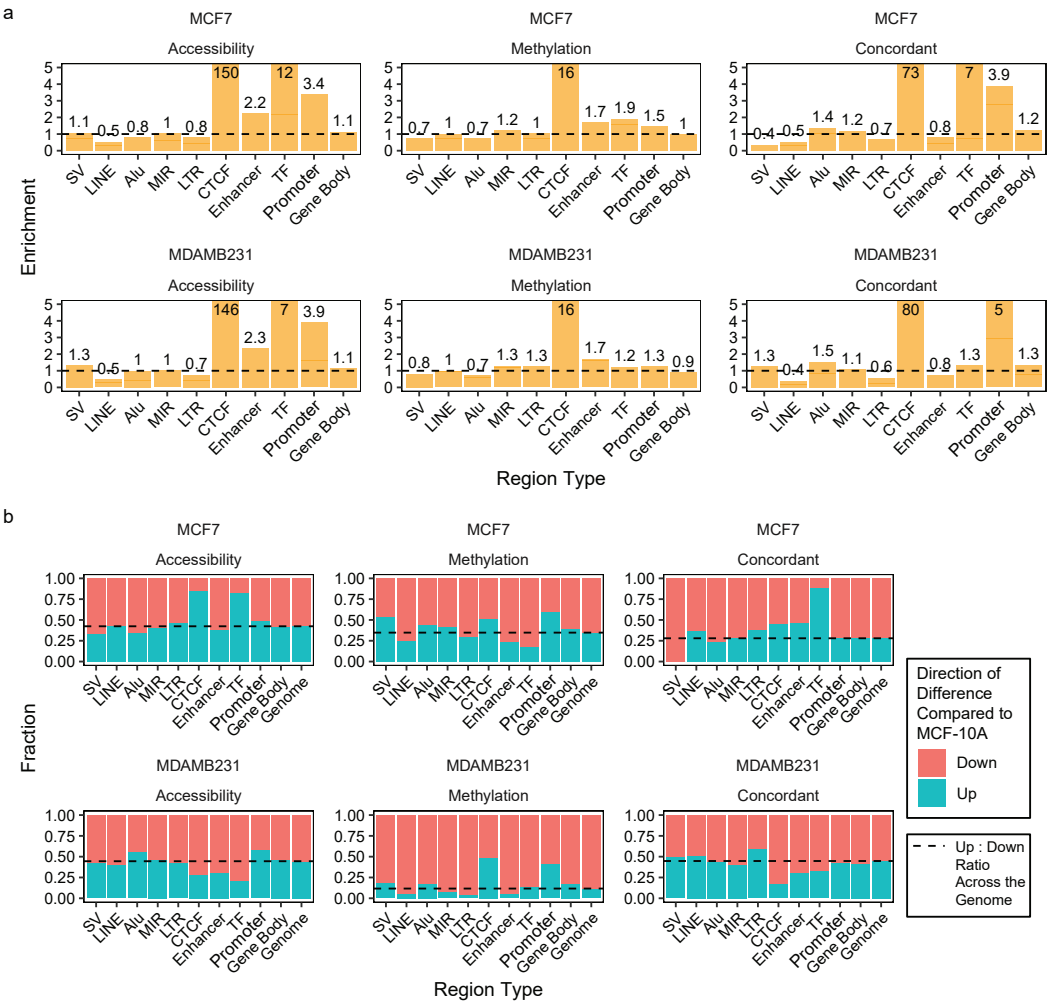


Figure 2.18: Enrichment of differential epigenetic regions in various genomic contexts. (a) Enrichment of DMRs, DARs, and concordantly differential regions are calculated for each genomic context against the abundance across the genome. **(b)** The makeup of directions of the differential regions for each of these genomic contexts presented as fractions, compared to the makeup of the direction across the whole genome (dotted line).

18,955 SVs across all three breast lines and compared these using SURVIVOR (**Supplementary Table 2.5**) (Sedlazeck et al., 2018; Jeffares et al., 2017). We found that while many of the insertions were homologous to known repetitive elements (18% of MCF-10A, 30% of MCF-7, and 25% of MDA-MB-231), the

repetitive sequences were shared among the three, indicating that there is no difference in repetitive element activities between the three cell lines. Consistent with this finding, we observed no enrichment of differential epigenetic regions in repetitive elements (**Figure** 2.18). The majority of the SVs were singletons (65.9%), and 1,805 SVs occurred in both of the cancer subtypes and not in MCF10A. While DMRs and DARs were not enriched in regions surrounding SVs (**Figure** 2.18), we were able to identify SVs that occurred only on one cell line and coincide with differential epigenetic states, demonstrating the ability of nanoNOMe to evaluate the epigenome in and around SVs (**Supplementary Figure** 2.23). As an example, we found an insertion on chr6:169,976,00 that occurred on both MCF-7 and MDA-MB-231 but not in MCF-10A, which also showed a region 1kb downstream of the insertion that was hypermethylated and less accessible (**Figure** 2.19). These changes in the SV-containing cancer subtypes show that the structural variants in the region may have caused epigenetic changes that made the region epigenetically inactive

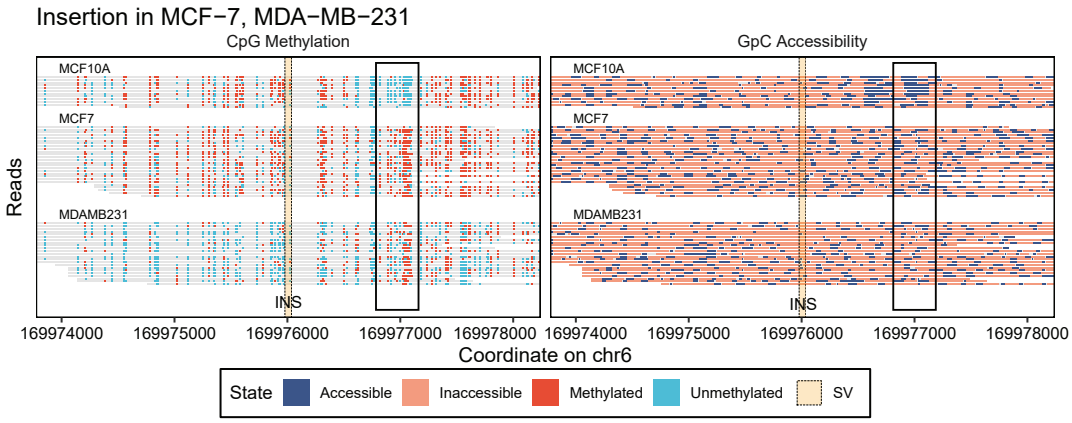


Figure 2.19: Epigenetic differences on an insertion only present in MCF-7 and MDA-MB-231. Methylation and accessibility of individual reads near an insertion that is present in MCF-7 and MDA-MB-231 but not in MCF-10A, showing changes in methylation and accessibility downstream of the insertion

2.4 Discussion

We have leveraged single molecule nanopore sequencing to directly examine endogenous CpG methylation and chromatin accessibility on long fragments of DNA. Leveraging long reads, we measured epigenetic states at genomic elements that were previously difficult to characterize, including repetitive elements and structural variations. We can also detect structural variations (SVs) with long reads, which are difficult to detect with short-read sequencing, and examine the epigenome in and around these SVs. With the ability to sequence parts of the genome that were previously difficult to sequence without sequence context-dependent bias, this method will serve as a valuable tool in furthering the role of DNA methylation and chromatin accessibility in regulation of genomic elements, or vice versa.

2.5 Methods

2.5.1 Methylation training and testing set generation

Genomic DNA from E. coli K12 MG1655 (ATCC 700926DQ) and NA12878, i.e. genomic DNA from GM12878 lymphoblast cell line (Coriell Institute), were first sheared to an average fragment size of 8 kb using Covaris g-tube shearing device (Covaris Cat. 520079). The fragmented DNA was PCR amplified to generate unmethylated DNA using the first steps of low input ligation kit SQK-LWP001 (ONT). Samples were end-repaired, deoxyadenosine(dA)-tailed, and ligated to amplification adaptors, followed by 11 cycles of PCR amplification. The resulting unmethylated, sheared DNA was methylated

with M. SssI (NEB Cat. M0226) for CpG methylation or M. CviPI (NEB Cat. M0227) for GpC methylation, or both enzymes for CpG+GpC methylation. Two cycles of 4-hour methylation were performed for each sample, and for each cycle of treatment the enzyme and methyl donor (S-adenosylmethionine) were replenished at the 2 hour mark.

2.5.2 Validation of DNA methylation by bisulfite sequencing

Near-complete methylation in the training samples (E. coli) and testing samples (GM12878) were validated by performing whole genome bisulfite sequencing on the Illumina MiSeq platform. NEBNext Ultra library preparation kit (NEB Cat. E7370) and Zymo EZ DNA methylation-lightning kit (Zymo Cat. D5030) were used to generate the bisulfite sequencing libraries. DNA from each sample was shared to 300 bp fragments using Bioruptor Pico (Diagenode), followed by end-repair and dA-tailing. Methylated universal adaptor (NEB Cat. E7535) was ligated using the Blunt/TA ligase from the kit. The adaptor-ligated samples were bisulfite-converted, quenched, and cleaned-up before PCR amplification with multiplexing primers and uracil-tolerant Taq polymerase (KAPA HiFi Uracil+ (Roche Cat. KK2801)). The resulting DNA sequencing library was sequenced on an Illumina MiSeq device using V2 300-cycle chemistry.

2.5.3 Processing of bisulfite sequencing data

The resulting fastq files were preprocessed by removing adaptor sequences and trimming low quality 3′ ends using Trim Galore version 0.6.3 (https:

//github.com/FelixKrueger/TrimGalore) with default parameters. Then, data was analyzed using Bismark version 0.19.0 (Krueger and Andrews, 2011). After alignment, PCR duplicates were removed using Picard tools MarkDuplicates module version 2.20.2 (http://broadinstitute.github.io/picard/). Reads were truncated at the 3′ end by 2 bases at the 5′ end and 1 base at 3′ end to minimize methylation bias at the ends of reads introduced during the library preparation. The total number of methylated cytosine residues and unmethylated cytosine residues were counted to calculate methylation percentages of the samples.

2.5.4 nanopolish methylation training for dual CpG/GpC methylation calling

To train the methylation calling models we generated nanopore sequencing data for E. coli gDNA treated with M.SssI (to methylate CpGs), M.CviPI (to methylate GpCs), and both M.SssI and M.CviPI (to methylate in both contexts) (described above). The three datasets were basecalled with Guppy (version 3.0.3) and aligned to the E. coli genome using NGMLR version 0.2.8. The reference genomes for each dataset were then modified by converting Cs to Ms in the appropriate context. We then merged the three reference genomes and three BAM files together and downsampled the alignments to 10% coverage to reduce model training time in the subsequent step. At the end of this preprocessing, we had a dataset with a mixture of reads that have CpG methylation, GpC methylation, or both, and matching reference sequences to align each read to indicate the pattern of methylation in each read.

The k-mer states for the CpG/GpC model were trained using âĂIJnanopolish trainâĂİ module in nanopolish cpggpc_new_train branch (commit c409580). Model training was run for 10 iterations and the final model was used for the subsequent methylation calling.

2.5.5 Cell Culture

GM12878 lymphoblast cells were obtained from Coriell Institute and MCF-10A, MCF-7, and MDA-MB-231 breast cells were obtained from ATCC. GM12878 were grown in RPMI 1640 medium (Gibco Cat. 11875119) supplemented with 15% fetal bovine serum (FBS, Gibco Cat. 26140079) and 1% penicillin-streptomycin (P/S, Gibco Cat. 15140122). MCF-10A were grown in DMEM F-12 medium (Gibco Cat. 11320033) supplemented with 5% horse serum (Gibco Cat. 16050122), 10 µg/mL human insulin (Sigma Aldrich Cat. 19278), 20 ng/mL hEGF (Gibco Cat. PHG0311L), 100 ng/mL Cholera toxin (Sigma Aldrich Cat. C8052), 0.5 µg/mL Hydrocortisone (Sigma Aldrich Cat. H0135), and 1% P/S. MCF-7 and MDA-MB-231 were grown in DMEM (Gibco Cat. 11965118) supplemented with 10% FBS and 1% P/S.

2.5.6 Nucleosome footprinting via GpC methyltransferase

NOMe-seq was performed on the cells with adjustments for nanopore sequencing. Cells were collected by trypsinization, then nuclei were extracted by incubating in resuspension buffer (100 mM Tris-Cl, pH 7.4, 100 mM NaCl, 30 mM $MgCl_2$) with 0.25 % NP-40 for 5 minutes on ice. Intact nuclei were collected by centrifugation for 5 minutes at 500xg at 4 °C. Nuclei were subjected to a methylation labeling reaction using a solution of 1x M. CviPI Reaction

Buffer (NEB), 300 mM sucrose, 96 µM S-adenosylmethionine (SAM; New England Biolabs, NEB), and 200 U M. CviPI (NEB) in 500 µL volume per 500,000 nuclei. The reaction mixture was incubated at 37 °C with shaking on a thermomixer at 1,000 rpm for 15 minutes. SAM was replenished at 96 µM at 7.5 minutes into the reaction. The reaction was stopped by the addition of an equal volume of stop solution (20 mM Tris-Cl, pH 7.9, 600 mM NaCl, 1% SDS, 10 mM disodium EDTA). Samples were treated with proteinase K (NEB) at 55 °C for > 2 hours, and DNA was extracted via phenol:chloroform extraction and ethanol precipitation. After proteinase K treatment, and in all following steps, samples were handled with care using large orifice pipette tips to avoid excessive fragmentation of DNA.

2.5.7 Nanopore sequencing

Purified gDNA was prepared for nanopore sequencing following the protocol in the genomic sequencing by ligation kit LSK-SQK108 (ONT). Samples were first sheared to 10 kb using G-tubes (Covaris): by centrifuging 2-3µg of unfragmented gDNA at 5,000X G for 1 minute, then inverting the tube and centrifuging again. We sheared the DNA to 10 kb because it produces long fragments of DNA while maximizing the yield of nanopore sequencing. Shearing to larger sizes or unsheared DNA may be used to maximize the length of sequenced reads, with the caveat that sequencing yield will drop. In two samples (GM12878 samples 8 and 9), we targeted 20kb fragments, with an additional step of removing short fragments using the Short Read Eliminator module by Circulomics, following the manufacturerâĂŹs specifications. The

sheared samples were end-repaired and dA-tailed using NEBnext Ultra II end-repair module (NEB), followed by clean-up using 1X v/v AMPure XP beads (Beckman Coulter). Sequencing adaptors, comprised of leader adaptor DNA and motor proteins, were ligated to the end-prepared DNA fragments using Blunt/TA Ligase Master Mix (NEB), followed by clean-up using 0.4X v/v AMPure XP beads and sequencing kit reagents. >400 ng of adaptor ligated samples per flow cell were loaded onto FLO-MIN106 or PRO-002 flowcells and run on MinION Mk1b, GridION, or PromethION sequencers for up to 72 hours.

2.5.8 Data Processing (basecalling, alignment, and structural variant calling)

Raw current signals were converted to DNA sequences using Guppy version 3.0.3 (ONT), using the "high-accuracy" basecalling model (Jain et al., 2018). DNA sequences were aligned to hg38 human reference genome without alternative contigs using NGMLR version 0.2.8 with default settings for aligning Oxford nanopore reads (-x ont) (Sedlazeck et al., 2018). We used Sniffles version 1.0.11 (Sedlazeck et al., 2018) with default parameters to detect SVs across each sample and SURVIVOR version 1.0.734 to obtain a multi-sample VCF file.

2.5.9 Nanopolish methylation calling for dual CpG/GpC methylation

We modified the methylation calling module of nanopolish to be able to call methylation in multiple motifs simultaneously (github branch cpggpc_new_train,

commit c409580). As in our previous work, we start by grouping nearby CpG and GpC sites together (minimum distance of 5 to separate sites). We then calculate a likelihood for combinations of the grouped sites being methylated or unmethylated (either no sites methylated, all CpGs methylated, all GpCs methylated, or all sites both contexts methylated), using the k-mer states trained in the previous section, with the hidden Markov model we previously described. We then calculate a log-likelihood ratio for each motif (CpG, GpC), by summing the likelihoods across all sequences where the motif is methylated, or unmethylated.

2.5.10 Comparison of nanoNOMe with conventional methodologies

Bulk NGS methodologies comparable to nanoNOMe on GM12878 were used to compare and validate nanoNOMe. Whole genome bisulfite sequencing methylation frequencies were obtained from Encode accession ENCFF835NTC, normalized MNase-seq signals were obtained from Encode accession ENCSR000CXP, and normalized DNAse-seq signals were obtained from Encode accession ENCSR000EJD (ENCODE Project Consortium, 2012). ATAC-seq data was obtained from GEO accession GSE47753 and processed using the standard ENCODE pipeline (Buenrostro et al., 2013; ENCODE Project Consortium, 2012). Nanopore whole genome sequencing data was obtained from ENA accession PRJEB23027, and processed the same way as nanoNOMe (Jain et al., 2018).

For comparing mappability between WGBS, nanopore WGS, and nanoNOMe, the numbers of reads aligning to 200 bp bins of the genome were calculated.

GC-bias of the coverages were determined by calculating the percentages of C/G for each of the 200 bp bins and plotting the per-bin coverage against the CG percentage. To compare mappability in specific genomic contexts, a region was considered to be robustly mapped in a dataset if its coverage was between the 5th and 95th percentile of the genome-wide binned coverage. The upper threshold takes into account aberrantly highly mapped regions, while the lower threshold removes low mappability regions.

For comparison of nanoNOMe signals with conventional bulk methods, average methylation was calculated for each CpG and GpC site. To compare nanoNOMe CpG methylation to WGBS methylation, methylation frequencies for each CpG locus across the genome were compared pairwise between the two methods. To compare nanoNOMe GpC accessibility signal to normalized ATAC-seq and DNAse-seq signals, the intersections were determined from accessibility peaks of nanoNOMe, ATAC-seq, and DNAse-seq.

2.5.11 Metaplot Analysis

Methylation frequencies from WGBS and normalized MNase-seq signals at regions surrounding genomic features of interest (CTCF, TSS with respect to expression and histone modifications) were extracted for the generation of the metaplots. For each genomic feature, average methylation frequency and accessibility were aggregated with respect to distance from the feature, followed by taking the rolling average with a window of 50 bp. Known TSS and CGI were obtained from Gencode (release v29). TSSs were grouped by expression quartile based on RNA-seq of GM12878 (ENCODE accession

ENCSR843RJV), and by the presence of ChIP-seq peaks of histone modifications H3K4me3 (ENCODE accession ENCSR057BWO) and H3K27me3 (ENCODE accession ENCSR000AKD) within 1kb of the TSS. CTCF binding sites were determined by overlapping computationally predicted CTCF binding sites with conservative IDR peaks in ChIP-seq of CTCF on GM12878 (ENCODE accession ENCSR000AKB) and removing peaks that fell within 2kb of known TSS (Ziebarth, Bhattacharya, and Cui, 2013).

2.5.12 Enrichment analysis of differential epigenetic regions on genomic contexts

To calculate the enrichment of DMRs, DARs, and concordantly differential regions in various genomic contexts, we first calculated the total width of the genome and the total width of the genomic contexts of interest that contain CpG and GpC data. Then the total number of differential regions was divided by the total width of the genome, which is the expected abundance of differential regions. This was used as the baseline against the total numbers of differential regions in genomic contexts of interest divided by the total widths of the genomic contexts to generate the final values of enrichment. For TSS and small TF binding sites, we used 1000 bp regions centered on the genomic elements.

2.6 Acknowledgments

Funding: This study was supported by National Human Genome Research Institute (NHGRI project 5R01HG009190)

Competing Interests: WT has two patents (8,748,091 and 8,394,584) licensed to Oxford Nanopore. JTS received research funding from ONT. IL, TG, NS, JTS, FJS and WT have received travel funds to speak at meetings from ONT.

Data Accessions: NanoNOMe data of GM12878, MCF-10A, MCF-7, and MDA-MB-231 are available at NCBI Bioproject ID PRJNA510783 (http://www.ncbi.nlm.nih.gov/bioproject/510783). Source code is available at https://github.com/timplab/nanoNOMe.

2.7 Supplementary Material

Sample	Methylation	Total Reads	Reads with MAPQ >= 20	Fraction	Yield (Gb)	Coverage
E. coli	Unmethylated	724,887	707,199	0.98	4.14	891.24
E. coli	CpG	1,396,674	1,355,969	0.97	7.39	1592.70
E. coli	GpC	1,251,301	1,209,030	0.97	7.15	1540.85
E. coli	CpGGpC	1,084,568	1,048,904	0.97	6.41	1381.27
NA12878	Unmethylated	2,999,723	2,803,707	0.93	10.28	3.18
NA12878	CpG	2,951,563	2,748,099	0.93	8.47	2.62
NA12878	GpC	2,941,260	2,723,240	0.93	9.28	2.87
NA12878	CpGGpC	2,848,032	2,631,771	0.92	9.14	2.83

Table 2.2: Nanopore sequencing yields of testing and training samples. After sequencing the training and testing sets on a minION nanopore sequencing, total numbers of methylated loci and unmethylated loci for each sample was tabulated to calculate the percent methylation per sample.

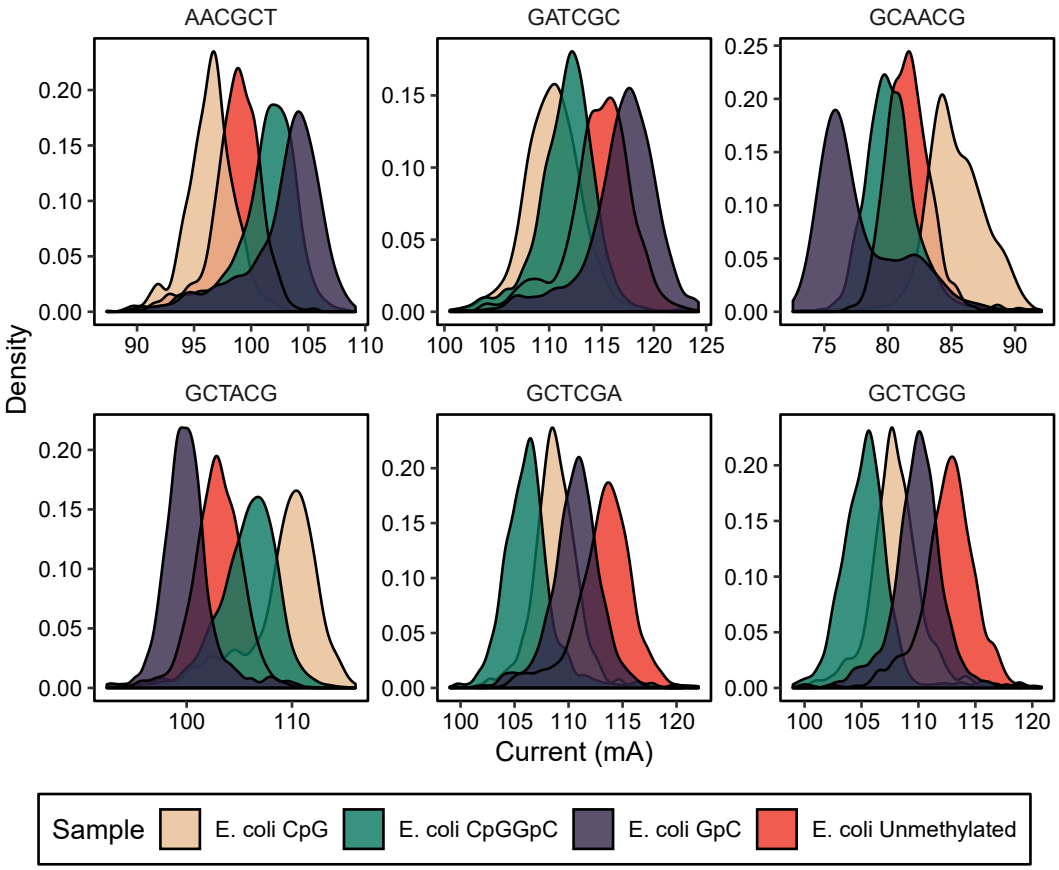


Figure 2.20: Distributions of current modulation in select 6-mers.

		Genomic Contexts				
Motif	Metric	CGI	Genes	LINE	Promoters	SINE
CG	Accuracy	0.99	1.00	1.00	1.00	1.00
CG	Call Rate	1.09	1.01	1.01	1.04	1.01
GC	Accuracy	0.99	1.00	1.00	1.00	1.00
GC	Call Rate	0.97	1.01	1.01	1.00	1.01

Table 2.3: Relative accuracy and call rates for notable genomic contexts. Accuracy and call rate with respect to genomic contexts in comparison to overall accuracy and call rate (context accuracy/call rate divided by overall accuracy/call rate). A relative value of 1 represents the same value in the context in comparison to overall value; No particular decrease in accuracy or call rate was observed based on genomic context.

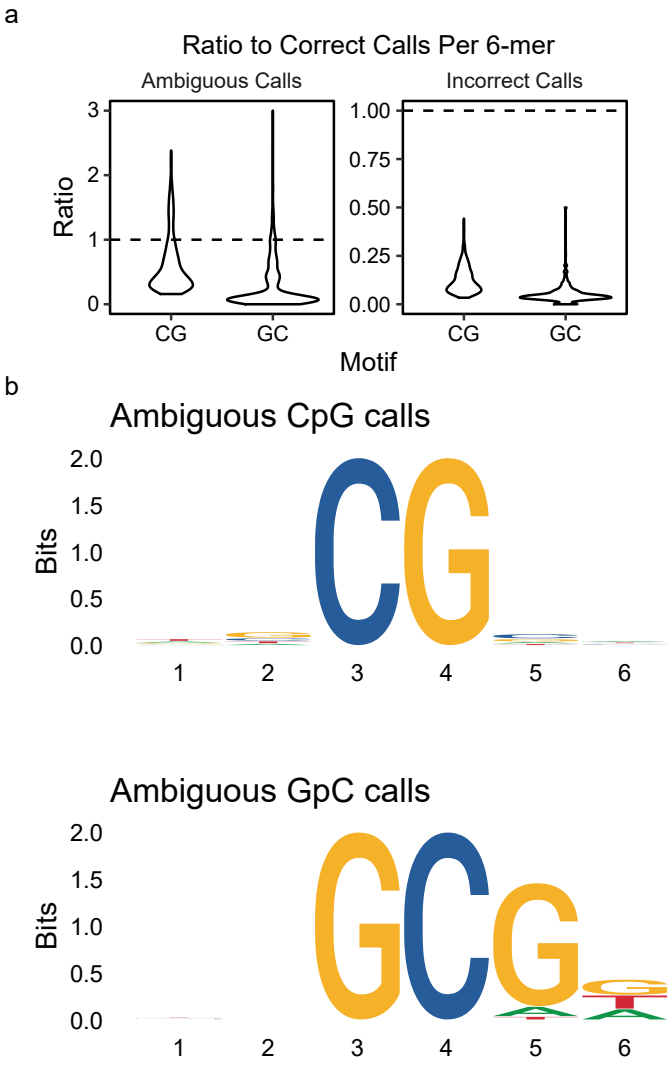


Figure 2.21: Sequence context dependence of methylation calling accuracy. (a) The ratio of ambiguous and incorrectly called calls to correctly called calls for every 6-mer with the methylation motif in the center (3th position for CpG and 4th position for GpC), and **(b)** motif analysis of enriched ambiguous 6-mers from (d) in (left) CpG calls and (right) GpC calls, showing enrichment in GCG motifs in GC calling, which are removed in our pipeline.

Cell	Experiment number	Flowcell type	Number of flowcells	Number of raw reads (M)	Total raw bases (Gb)	Aligned reads (M)	Aligned bases (Gb)	N50 length
GM12878	1	FLO-MIN106	2	1.64	12.95	1.30	11.20	11,475
GM12878	2	FLO-MIN106	2	1.90	14.86	1.56	12.68	10,736
GM12878	3	FLO-MIN106	2	0.81	6.53	0.68	5.66	10,526
GM12878	4	FLO-MIN106	2	1.03	11.10	0.89	9.79	17,346
GM12878	5	FLO-MIN106	2	1.95	14.77	1.55	12.11	11,635
GM12878	6	FLO-MIN106	2	2.09	15.42	1.69	13.55	11,156
GM12878	7	FLO-PROM002	1	11.51	72.84	8.90	62.62	9,791
GM12878	8	FLO-PROM002	1	5.40	85.51	4.80	74.94	20,850
GM12878	9	FLO-PROM002	1	5.66	64.28	5.03	54.37	20,626
MCF10A	1	FLO-MIN106	2	0.86	6.52	0.60	5.70	11,428
MCF10A	2	FLO-MIN106	3	3.59	33.11	3.05	29.64	11,888
MCF10A	3	FLO-MIN106	4	4.96	41.96	4.09	37.06	11,215
MCF7	1	FLO-MIN106	1	1.17	9.71	1.03	8.77	11,210
MCF7	2	FLO-MIN106	2	0.78	5.29	0.52	4.72	11,652
MCF7	3	FLO-MIN106	5	5.42	44.25	4.48	39.61	12,462
MCF7	4	FLO-MIN106	3	1.62	17.51	1.46	16.05	18,532
MDAMB231	1	FLO-MIN106	1	0.89	8.22	0.78	7.39	11,531
MDAMB231	2	FLO-MIN106	2	2.10	19.01	1.82	17.30	11,488
MDAMB231	3	FLO-MIN106	3	3.09	33.06	2.68	30.14	14,583
MDAMB231	4	FLO-MIN106	3	1.87	22.10	1.67	20.06	15,783

Table 2.4: Individual nanopore sequencing run metrics of nanoNOMe samples. Nanopore sequencing run statistics for individual sequencing experiments performed, before pooling them by cell line to achieve the final yields.

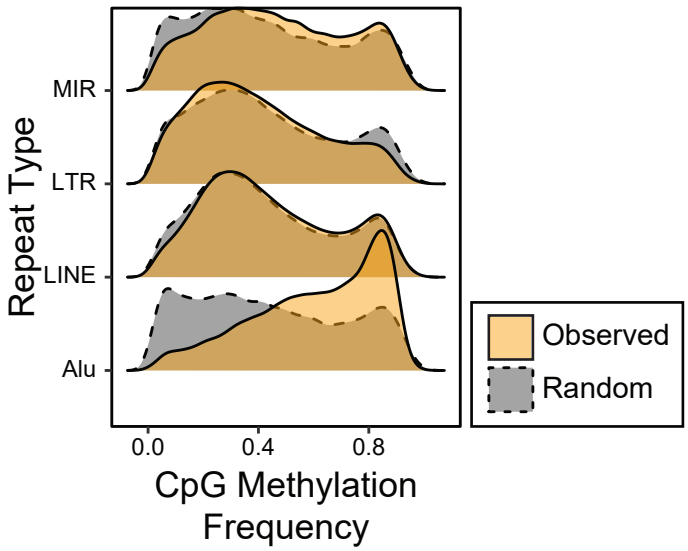


Figure 2.22: Per-Region nanoNOMe frequency in repetitive elements. Distribution of observed per-region average methylation frequency in repetitive elements in comparison to random regions across the genome of the same lengths.

Sample	DEL	INS	DUP	INV	TRA	Total
MCF10A	2245	1381	188	37	57	3908
MCF7	2297	1139	225	63	110	3834
MDAMB231	2605	1754	305	44	37	4745
MCF7 + MDAMB231	1020	669	92	9	15	1805
MCF10A + MDAMB231	1562	1124	156	14	11	2867
MCF10A + MCF7	1032	655	69	19	21	1796
Total	10761	6722	1035	186	251	18955

Inclusive counts

MCF10A	4839	3160	413	70	89	8571
MCF7	4349	2463	386	91	146	7435
MDAMB231	5187	3547	553	67	63	9417

Table 2.5: Summary of structural variations detected in breast cell lines. Structural variations types are deletions (DEL), translocations (TRA), duplications (DUP), inversions (INV), and insertions (INS), and are grouped by uniquely occurring (first three lines), commonly occurring in any combination of two cell lines (second three lines), and a summary of inclusive counts for each cell line. SVs of < 50bp were filtered out.

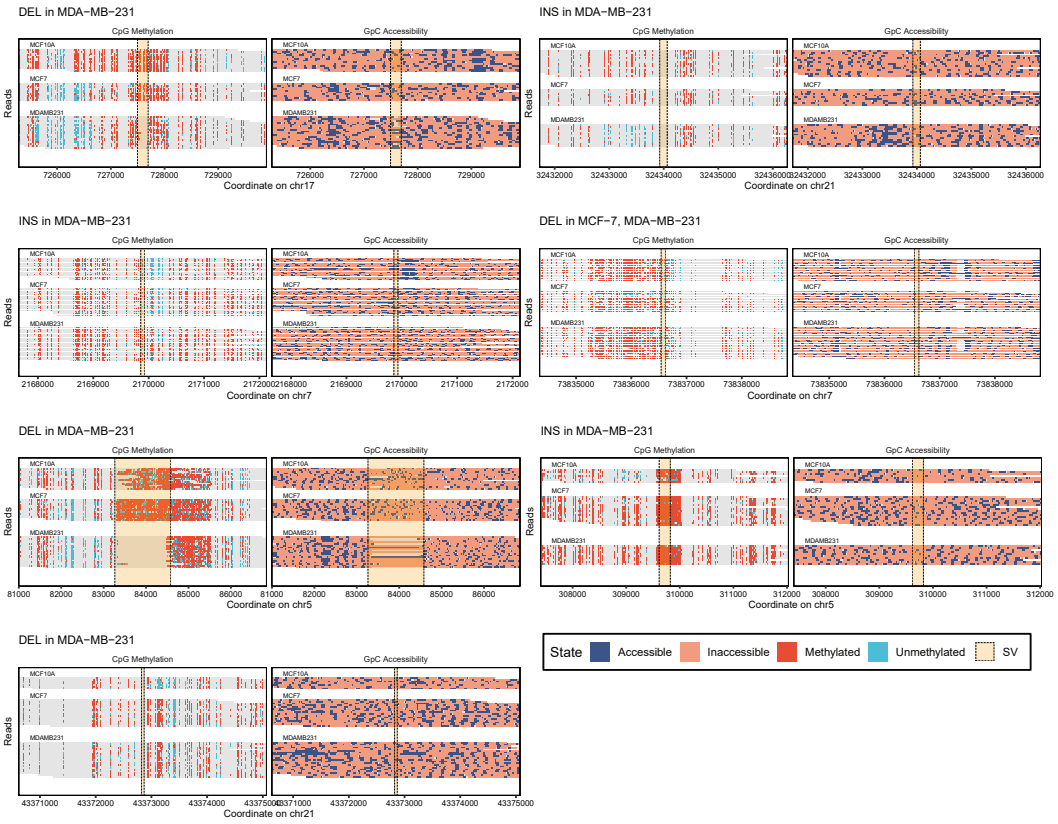


Figure 2.23: Structural variations and differential epigenetics. Single-read methylation and accessibility plots of regions that had an SV in the cancer subtypes and not in MCF-10A, as well as differential methylation and accessibility in the cancer subtypes in comparison to MCF-10A.

References

Frommer, M, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, and C L Paul (1992). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 89.5, pp. 1827–1831.

Krueger, Felix and Simon R Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". en. In: *Bioinformatics* 27.11, pp. 1571–1572.

Hansen, Kasper D, Benjamin Langmead, and Rafael A Irizarry (2012). "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions". en. In: *Genome Biol.* 13.10, R83.

Boyle, Alan P, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford (2008). "High-resolution mapping and characterization of open chromatin across the genome". en. In: *Cell* 132.2, pp. 311–322.

Buenrostro, Jason D, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf (2015). "Single-cell chromatin accessibility reveals principles of regulatory variation". en. In: *Nature* 523.7561, pp. 486–490.

Henikoff, Jorja G, Jason A Belsky, Kristina Krassovsky, David M MacAlpine, and Steven Henikoff (2011). "Epigenome characterization at single base-pair resolution". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.45, pp. 18318–18323.

Kelly, Theresa K, Yaping Liu, Fides D Lay, Gangning Liang, Benjamin P Berman, and Peter A Jones (2012). "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules". en. In: *Genome Res.* 22.12, pp. 2497–2506.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). "Detecting DNA cytosine methylation using nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 407–410.

Rand, Arthur C, Miten Jain, Jordan M Eizenga, Audrey Musselman-Brown, Hugh E Olsen, Mark Akeson, and Benedict Paten (2017). "Mapping DNA methylation with high-throughput nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 411–413.

Shipony, Zohar, Georgi K Marinov, Matthew P Swaffer, Nicholas A Sinnott-Armstrong, Jan M Skotheim, Anshul Kundaje, and William J Greenleaf (2020). "Long-range single-molecule mapping of chromatin accessibility in eukaryotes". en. In: *Nat. Methods* 17.3, pp. 319–327.

Wang, Yunhao, Anqi Wang, Zujun Liu, Andrew L Thurman, Linda S Powers, Meng Zou, Yue Zhao, Adam Hefel, Yunyi Li, Joseph Zabner, and Kin Fai Au (2019). "Single-molecule long-read sequencing reveals the chromatin basis of gene expression". en. In: *Genome Res.* 29.8, pp. 1329–1342.

Zook, Justin M, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, Elizabeth Henaff, Alexa B R McIntyre, Dhruva Chandramohan, Feng Chen, Erich Jaeger, Ali Moshrefi, Khoa Pham, William Stedman, Tiffany Liang, Michael Saghbini, Zeljko Dzakula, Alex Hastie, Han Cao, Gintaras Deikus, Eric Schadt, Robert Sebra, Ali Bashir, Rebecca M Truty, Christopher C Chang, Natali Gulbahce, Keyan Zhao, Srinka Ghosh, Fiona Hyland, Yutao Fu, Mark Chaisson, Chunlin Xiao, Jonathan Trow, Stephen T Sherry, Alexander W Zaranek, Madeleine Ball, Jason Bobe, Preston Estep, George M Church, Patrick Marks, Sofia Kyriazopoulou-Panagiotopoulou, Grace X Y Zheng, Michael Schnall-Levin, Heather S Ordonez, Patrice A Mudivarti, Kristina Giorda, Ying Sheng, Karoline Bjarnesdatter Rypdal, and Marc Salit (2016). "Extensive sequencing of seven human genomes to characterize benchmark reference materials". en. In: *Sci Data* 3, p. 160025.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414, pp. 57–74.

Eberle, Michael A, Epameinondas Fritzilas, Peter Krusche, Morten Kallberg, Benjamin L Moore, Mitchell A Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J Humphray, Aaron L Halpern, Semyon Kruglyak, Elliott H Margulies, Gil McVean, and David R Bentley (2016). "A reference dataset of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". en.

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan,

Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". en. In: *Nat. Biotechnol.* 36.4, pp. 338–345.

Olova, Nelly, Felix Krueger, Simon Andrews, David Oxley, Rebecca V Berrens, Miguel R Branco, and Wolf Reik (2018). "Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data". en. In: *Genome Biol.* 19.1, p. 33.

Ji, Lexiang, Takahiko Sasaki, Xiaoxiao Sun, Ping Ma, Zachary A Lewis, and Robert J Schmitz (2014). "Methylated DNA is over-represented in whole-genome bisulfite sequencing data". en. In: *Front. Genet.* 5, p. 341.

Lander, E S and M S Waterman (1988). "Genomic mapping by fingerprinting random clones: a mathematical analysis". en. In: *Genomics* 2.3, pp. 231–239.

Bell, A C and G Felsenfeld (2000). "Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene". en. In: *Nature* 405.6785, pp. 482–485.

Radman-Livaja, Marta and Oliver J Rando (2010). "Nucleosome positioning: how is it established, and why does it matter?" en. In: *Dev. Biol.* 339.2, pp. 258–266.

Eckhardt, Florian, Joern Lewin, Rene Cortese, Vardhman K Rakyan, John Attwood, Matthias Burger, John Burton, Tony V Cox, Rob Davies, Thomas A Down, Carolina Haefliger, Roger Horton, Kevin Howe, David K Jackson, Jan Kunde, Christoph Koenig, Jennifer Liddle, David Niblett, Thomas Otto, Roger Pettett, Stefanie Seemann, Christian Thompson, Tony West, Jane Rogers, Alex Olek, Kurt Berlin, and Stephan Beck (2006). "DNA methylation profiling of human chromosomes 6, 20 and 22". en. In: *Nat. Genet.* 38.12, pp. 1378–1385.

Holliday, Deborah L and Valerie Speirs (2011). "Choosing the right cell line for breast cancer research". en. In: *Breast Cancer Res.* 13.4, p. 215.

Messier, Terri L, Jonathan A R Gordon, Joseph R Boyd, Coralee E Tye, Gillian Browne, Janet L Stein, Jane B Lian, and Gary S Stein (2016). "Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes". en. In: *Oncotarget* 7.5, p. 5094.

Ehrlich, Melanie (2009). "DNA hypomethylation in cancer cells". en. In: *Epigenomics* 1.2, pp. 239–259.

Rothenberg, Ellen V (2014). "The chromatin landscape and transcription factors in T cell programming". en. In: *Trends Immunol.* 35.5, pp. 195–204.

Sedlazeck, Fritz J, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz (2018). "Accurate detection of complex structural variations using single-molecule sequencing". en. In: *Nat. Methods* 15.6, pp. 461–468.

Jeffares, Daniel C, Clemency Jolly, Mimoza Hoti, Doug Speed, Liam Shaw, Charalampos Rallis, Francois Balloux, Christophe Dessimoz, Jürg Bähler, and Fritz J Sedlazeck (2017). "Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast". en. In: *Nat. Commun.* 8, p. 14061.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". en. In: *Nat. Methods* 10.12, pp. 1213–1218.

Ziebarth, Jesse D, Anindya Bhattacharya, and Yan Cui (2013). "CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization". en. In: *Nucleic Acids Res.* 41.Database issue, pp. D188–94.

Chapter 3

Allele-specific and single-molecule epigenomic analysis

3.1 Abstract

We have shown the use of nanopore sequencing for studying the endogenous CpG methylation and exogenously labeled chromatin accessibility (nanoNOMe). In addition to the bulk analysis, there are several ways to utilize the long reads of nanopore sequencing. The long single-molecule resolution allows footprinting of protein and nucleosome binding, allowing the separation of reads based on protein binding state. On gene promoters, combinatorial epigenetic states can be measured on individual molecules. The long reads also make it possible to robustly assign reads to haplotypes, separating the data into maternal and paternal alleles. This also separates the methylation and accessibility information embedded in the reads, allowing allele-specific epigenetic analysis across the genome. We use existing SNV data on GM12878 to phase the nanoNOMe reads and present the first fully phased human epigenome, consisting of chromosome-level allele-specific profiles of CpG methylation

and chromatin accessibility.

3.2 Introduction

Previous studies have demonstrated that nucleosome positioning and DNA accessibility are heterogeneous even within a homogeneous cell population, highlighting the importance of probing these features on individual copies of the DNA (Buenrostro et al., 2015; Lai et al., 2018). This becomes even more important when the cell population is heterogeneous, such as in primary tissues or blood samples. This has led to adaptation of chromatin profiling and epigenetic assays to single-cell approaches (Guo et al., 2013; Smallwood et al., 2014; Clark et al., 2018; Lai and Pugh, 2017; Satpathy et al., 2019). Single-cell adaptations of bisulfite sequencing assays have characterized the heterogeneity of CpG methylation in embryonic stem cells and conversely the consistency of CpG methylation in haploid cells (Guo et al., 2013; Smallwood et al., 2014). Single-cell MNase-seq revealed that heterochromatic regions have large variation in nucleosome positioning while the nucleosomes are uniformly positioned, and euchromatic regions have lower variation of nucleosome positioning with inconsistent spacing especially at the hyperaccessible regions (Lai and Pugh, 2017). Application of single-cell ATAC-seq on tumor biopsies has identified a subset of cells that show response to anti-tumor treatments, demonstrating the potential of using single-cell approaches to detect subtle changes in response to stimuli (Satpathy et al., 2019). More recently, single-cell adaptation of NOMe-seq was coupled to single-cell RNA-seq, presenting measurements of multiple layers of gene regulation on single-cell resolution (Clark

et al., 2018). Nanopore sequencing is similar to these single-cell sequencing methodologies: the long reads can be utilized to link epigenetic patterns across long distances on individual reads, making the observations of these patterns on single-molecule resolution.

Normal somatic mammalian cells are diploid, meaning each cell carries two copies of each chromosome in the nucleus, one set (i.e. allele) from each parent. The function of epigenetics has been widely implicated in allele-specific activity of chromosomes, and it is well known that CpG methylation plays a pivotal role in X chromosome inactivation (Han, Lee, and Szabó, 2008; Fournier et al., 2002; Singer-Sam and Riggs, 1993). However, this difference in epigenetic signatures between alleles is problematic in NGS methodologies because the signal from one allele can reduce the signal from the other allele, and separating the alleles is not a trivial task, even in single-cell sequencing methodologies. This problem of ploidy becomes even more confounding in cases of aneuoploidy, where a cell has abnormally large number of chromosomes, which is widely prevalent in all types of cancers (Lengauer, Kinzler, and Vogelstein, 1998).

One way to resolve the parent-of-origin of DNA is by annotating single nucleotide variations (SNVs) in a parental trio : in an individual and both of his or her parents (1000 Genomes Project Consortium et al., 2010). By tracing heterozygous mutations from the individual to mutations present in the parents, the parental origin of the mutation can be deduced. Studies have used these haplotype annotations to phase DNA and RNA sequences and resolve allele-specific genomes and transcriptomes (Rozowsky et al., 2011;

Bansal and Bafna, 2008). However, to phase the DNA strands to an allele, a heterozygous SNV must be present within the read. Because of this, phasing NGS reads is limited to regions with high density of heterozygous SNVs. Nanopore sequencing, on the other hand, generates long reads, so each read has a greater chance of encountering one or more heterozygous SNPs which can be used to phase the reads into maternal or paternal origin.

Here we show the application of nanoNOMe in 1) observing chromatin accessibility and CpG methylation patterns on single-read resolution and 2) phasing the reads into their parental alleles, generating a first genome-wide map of allele-specific CpG methylation and chromatin accessibility on a human genome.

3.3 Results

3.3.1 Co-occurrence of accessibility patterns to observe cis-regulatory interactions

We explored the applicability of long reads generated from nanoNOMe in detecting patterns of the epigenetic features. Using the epigenetic features encoded on long sequences of reads, we can observe patterns of these features along the length of the reads, e.g. positioning of multiple nucleosomes on single strands of DNA by oscillation of GpC methylation. However, the inherent heterogeneity of the chromatin due to the dynamic nature of nucleosome positioning is directly translated to the single-read data, making it difficult to observe patterns of epigenetic features (Lai and Pugh, 2017). The biological heterogeneity of DNA accessibility is further compounded by errors

associated with the enzymatic methylation, such as imperfect methylation efficacy, non-specific methylation, and dissociation of nucleosomes in a small fraction of DNA during lysis. In order to resolve patterns of methylation and DNA accessibility on single-read resolution, we have to account for the heterogeneity and noise.

To that end, we focused on co-occurrences of methylated or unmethylated cytosine on each read, where the co-occurrence is defined by same type of event (methylated or unmethylated) being observed at two distinct positions on a given read (see **Methods** 3.5.1). Piling up the co-occurrence across reads in a given region, we found that patterns of read-level nucleosome positioning across the length of reads can be resolved using a co-occurrence matrix (**Figure** 3.1). Because this analysis measures the relationships of methylation - in the case of methylated co-occurrence - and unmethylation - i nthe case of unmethylated co-occurrence - between positions on individual reads, the peaks in the heatmap highlight locations of co-occurring nucleosome positioning and distances between them, whereas the average plot only indicates that nucleosomes were present without any relationship to other locations.

3.3.2 Improving single-molecule accessibility measurements using a smoothing estimator

Though the co-occurrence matrices have the ability to show patterns of accessibility that bulk measurements cannot, the method still relies on aggregation of the measurements to identify hotspots of interactions. So this is not a true single-molecule analysis, and will be unable to resolve interactions that occur in subgroups of the reads. To better utilize the epigenetic information

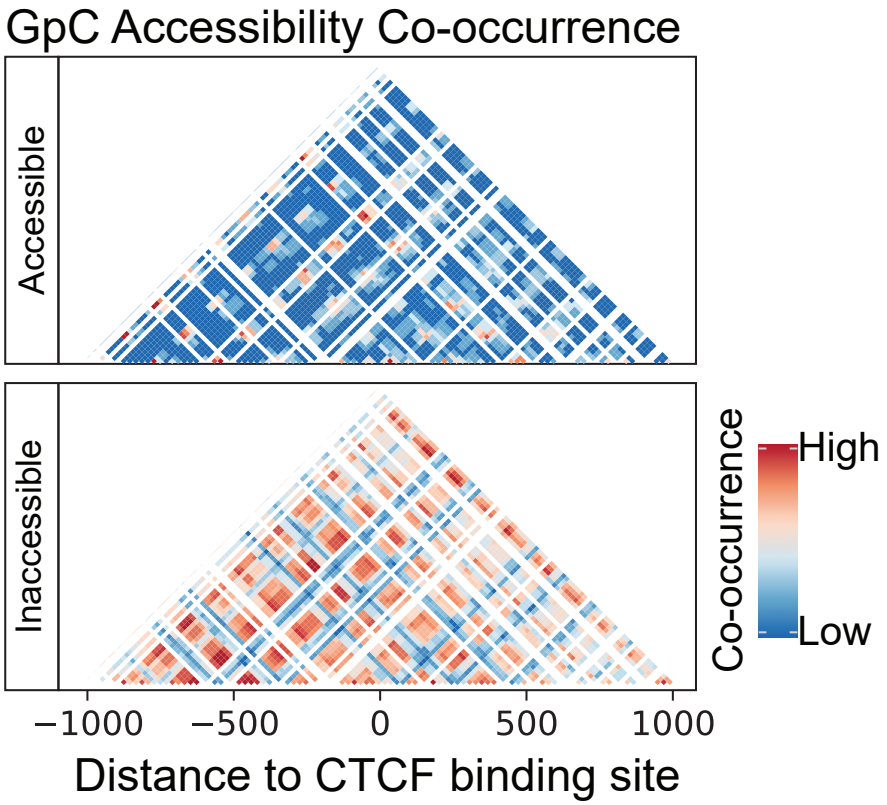


Figure 3.1: Matrix of accessibility co-occurrence on individual reads. Co-occurrence of accessibility and inaccessibility (GpC) are observed across the lengths of single reads and the co-occurrences events are piled up across all reads in the region resolves patterns such as co-associated nucleosome occupancy.

embedded on individual long strands of reads, we turned to direct ways to remove the noise in the single-read measurements of GpC accessibility. In the control testing samples, we examined the patterns of incorrect accessibility calls and found that 75% of incorrect calls were singletons, surrounded by correct calls, meaning that most of the noise is isolated. To remove the isolated noise, we implemented a method to estimate the accessibility of a given site using information from nearby GpC motifs on the same molecule, thereby dampening the isolated erroneous signal (see **Methods** 3.5.2). Briefly, we

applied a Gaussian kernel regression on the LLRs of accessibility calls using fixed genomic coordinate bandwidths and estimated accessibility across individual reads. We smoothed the GM12878 nanoNOMe data at CTCF binding sites, where accessibility profiles are the most consistent. We verified that the smoothing reduces the frequency of artifactual spikes in accessibility, evident by the removal of the very short lengths in the distribution plot of accessible and inaccessible runs (a run refers to a consecutive sequence of the same accessibility call) (**Figure** 3.2**a**). Aggregated frequencies of the smoothed calls still retained the oscillatory pattern, showing that the smoothing does not significanly decrease the ability to footprint nucleosome positioning (**Figure** 3.2**b**). Lastly, the smoothed calls allowed easier visualization of the patterns of accessibility on single-read level (**Figure** 3.2**c**).

3.3.3 Resolving regulatory protein binding on individual reads

We proceeded to characterize patterns of accessibility and methylation at CTCF binding sites on individual reads. First, we selected reads that span 2kb regions centered on 1,000 randomly selected CTCF binding sites from the 6,793 CTCF-binding sites with a ChIP-seq peak and another 1,000 from 4,288 binding sites without a ChIP-seq peak and examined runs of open and closed accessibility calls (**Figure** 3.3) (Ziebarth, Bhattacharya, and Cui, 2013; ENCODE Project Consortium, 2012). The position of closed runs differed between bound sites (sites with CTCF ChIP-seq peaks) and unbound sites (without peaks), with bound sites having a consistent pattern of nucleosome positioning. We found the length of closed runs corresponds to the units of

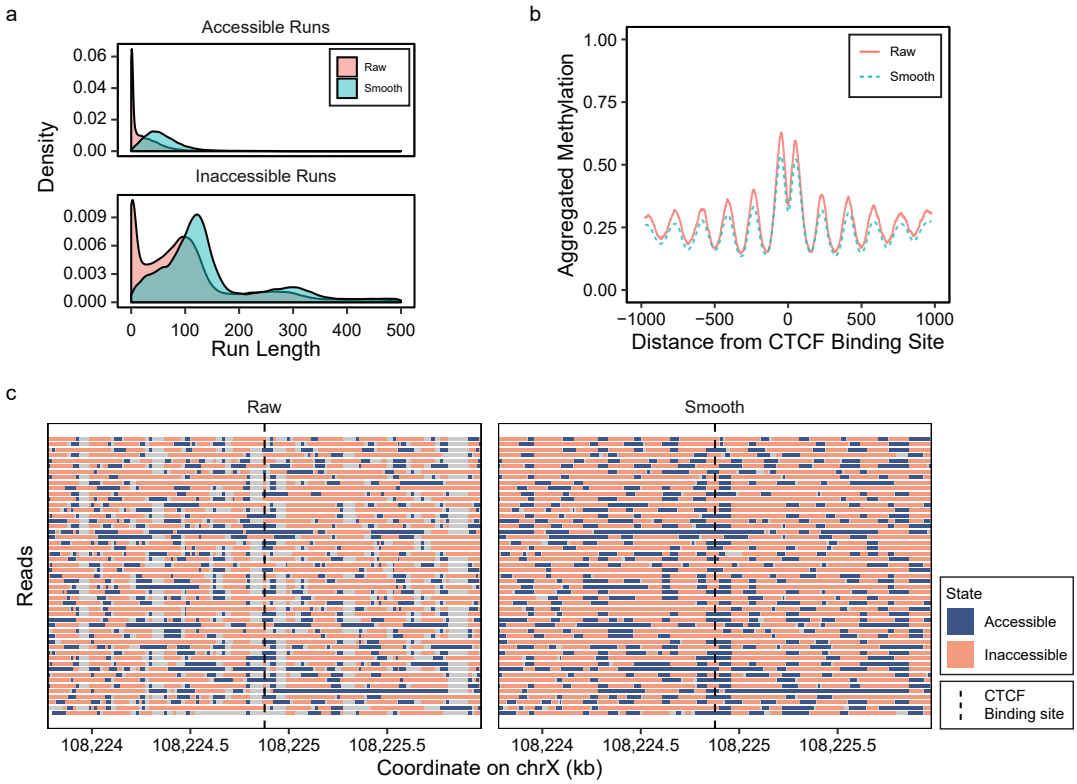


Figure 3.2: GpC accessibility kernel estimation on single reads. GpC methylation calls were smoothed using a Gaussian kernel estimator. **(a)** Distributions of length of open and closed runs and **(b)** metaplot of accessibility near CTCF binding sites before and after the smoothing, along with **(c)** example of read-level plot of accessibility from a 2kb region around a CTCF binding site.

nucleosomes, shown by hotspots of closed runs at 128 bp (mononucleosomes) and 310bp (dinucleosomes). Examining the length of the closed runs at the center of CTCF binding sites, we found a higher occurrence of shorter runs (<80bp), suggesting CTCF binding (**Supplementary Figure 3.17a**). This short length of inaccessibility by regulatory protein binding is consistent with previous findings of protein-DNA interactions via DNAse hypersensitivity and X-ray crystallography (Hesselberth et al., 2009; Luscombe et al., 2000; Boyle et al., 2008).

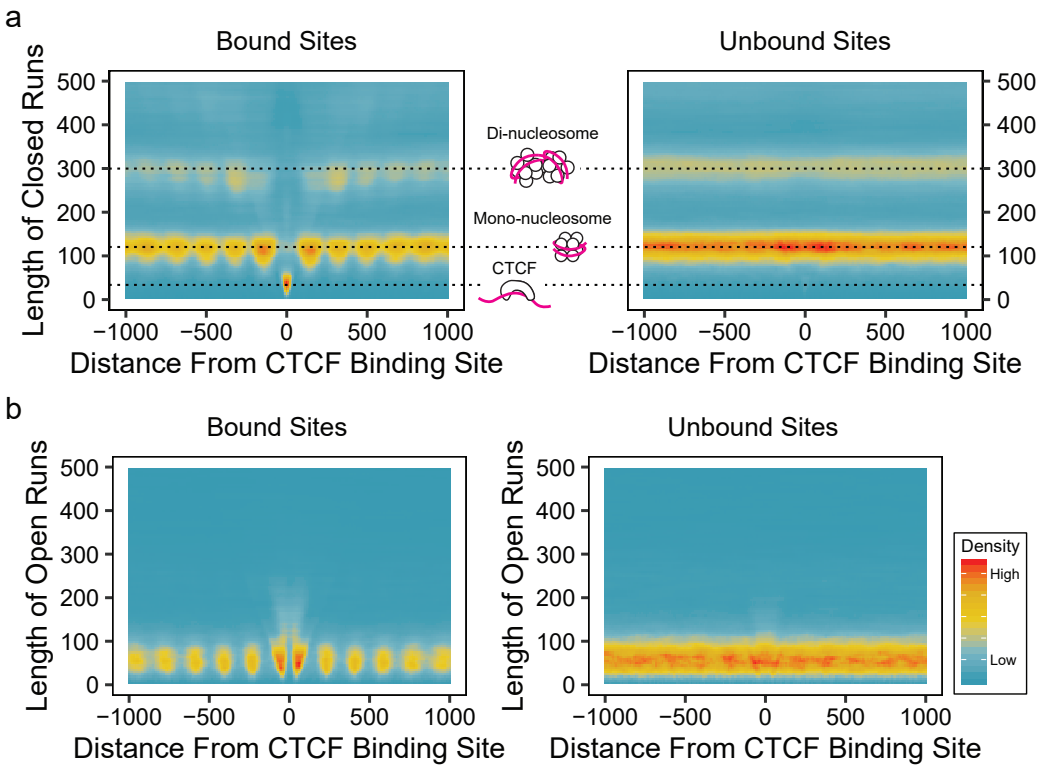


Figure 3.3: Single-molecule accessibility at CTCF binding regions. Heatmaps of lengths of **(a)** closed accessibility runs and **(a)** open accessibility runs on individual reads versus distance from CTCF binding sites, showing the relationship between accessibility run lengths and protein/nucleosome binding and the difference of these patterns on binding sites with and without ChIP-seq peaks.

Based on these observations, we used the length of closed runs at CTCF binding sites to infer CTCF binding state on individual reads (**Supplementary Figure** 3.17**b**, see **Methods** 3.5.3). We then separated reads based on their CTCF binding status and calculated the fraction of CTCF-bound reads for each site (**Figure** 3.4). Because nanoNOMe does not rely on enrichment or PCR to detect accessibility, the fraction of CTCF-bound reads represents a quantitative estimate of the degree of CTCF binding at the given site. We compared this data with CTCF Chip-seq peaks and found that the fraction of CTCF-bound

reads increased with increasing enrichment in ChIP-seq (Pearson correlation of 0.49), as opposed to sites outside of peaks that have consistently low fractions of CTCF-bound reads (median fraction of 0.02, **Figure** 3.5). Furthermore, we found the fraction of CTCF-bound reads varies widely even at places with ChIP-seq peaks.

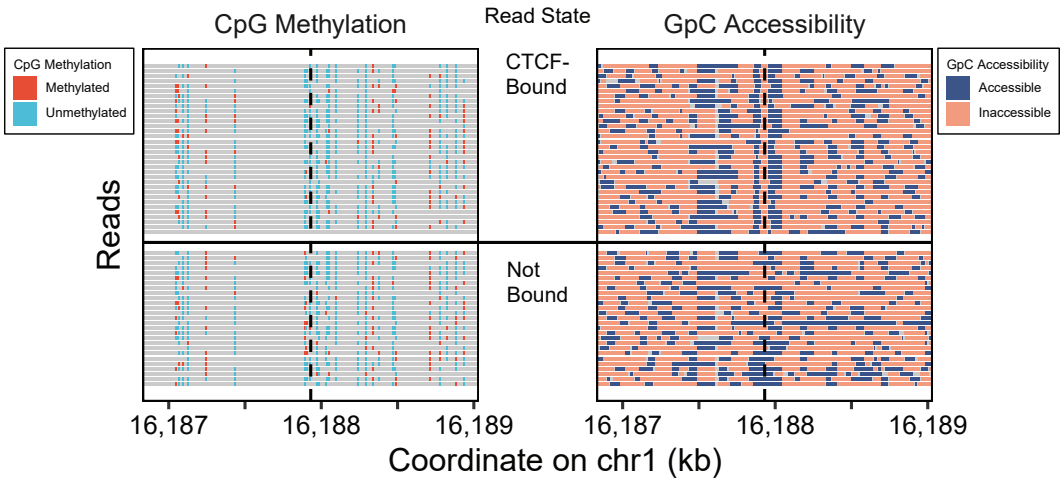


Figure 3.4: Per-read plot of methylation and accessibility on a CTCF binding site. CpG methylation and GpC accessibility of individual reads on a CTCF binding site, grouped by the predicted protein-binding state at the CTCF binding site.

To examine the global association of the affinity of the region to CTCF (presence of ChIP-seq peak) and the actual state of protein binding (read-level protein-binding prediction), we stratified the reads on CTCF binding sites based on the two layers of information : 1) whether the region had a peak in CTCF ChIP-seq and 2) whether we predicted the read to have protein binding at the binding site. We then generated methylation and accessibility metaplots of the aggregated profiles in each group (**Figure** 3.6). We found that reads on bound sites have consistently lower methylation even on reads that were not bound by CTCF. This agrees with a previous finding that demethylation of

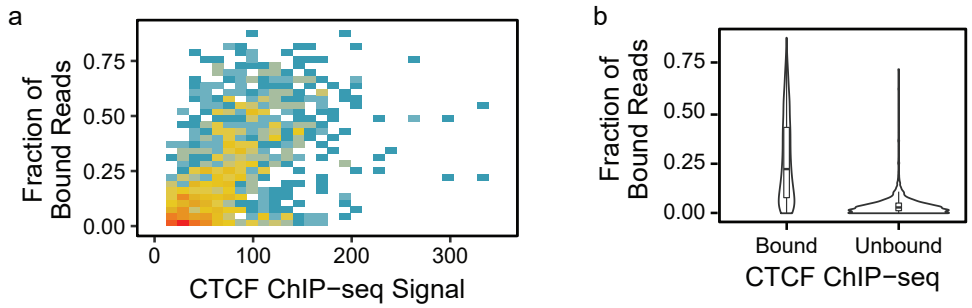


Figure 3.5: Comparison of protein-binding predictions with ChIP-seq signals. The read-level CTCF binding classification was tested by **(a)** comparing the per-site CTCF-bound read fractions with ChIP-seq coverage enrichment, showing that the ChIP-seq signal tends to increase with CTCF binding fraction, and **(b)** comparing the fractions in binding sites with ChIP-seq peaks to those without peaks, showing that sites with ChIP-seq peaks have higher fractions of CTCF binding.

CTCF binding sites increases the affinity of the protein to the binding site21. Similarly, we observed that nucleosomes are well-positioned irrespective of current CTCF occupancy.

3.3.4 Single-molecule combinatorial promoter epigenetic states

Next, we investigated epigenetic patterns on transcription start sites with single-read resolution on 1,000 randomly sampled genes from each expression quartile. On the TSS of highly transcribed genes, we observed a well-organized pattern of nucleosome positioning (low accessibility regions) and longer open runs representing nucleosome depleted regions (NDRs), whereas no pattern could be observed in lowly transcribed genes (**Figure** 3.7**a**). With decreasing expression, methylation increased and accessibility decreased around TSS (1kb for CpG and 200 bp for GpC) on a single-read level, in line with observations at the bulk level (**Figure** 3.7**b**). We used the methylation and accessibility

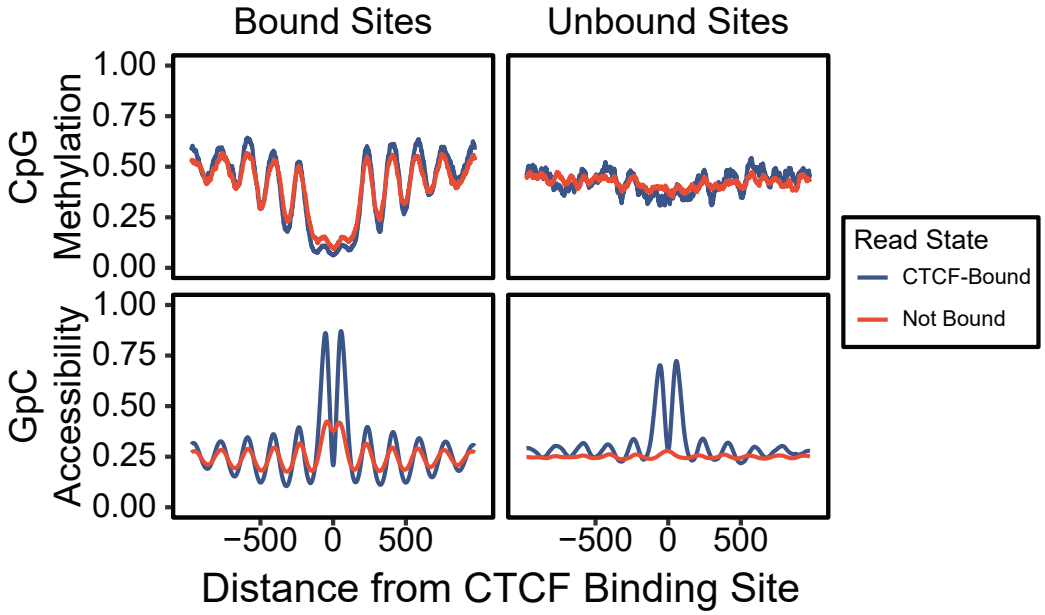


Figure 3.6: Metaplots of CTCF binding sites stratified by ChIP-seq peak and read-level protein binding. Metaplots at CTCF binding sites, separated by the presence of CTCF ChIP-seq peaks by panel and read-level CTCF binding by color, showing consistent epigenetic patterns on bound sites regardless of read-level binding state.

around the TSS to categorize reads into two groups (high and low frequency) for each feature (**Figure** 3.7**c**, see **Methods** 3.5.4). Mean CpG methylations for the two groups were 3% (demethylated) and 62% (methylated), and GpC groups had mean accessibilities of 20% (inaccessible) and 90% (accessible).

Combining the two features resulted in four possible combinatorial epigenetic states for each read (**Figure** 3.8). We observed that with increasing expression, fractions of concordantly active reads (low CpG methylation and high accessibility) increase and concordantly inactive (high CpG methylation and low accessibility) reads decrease (**Supplementary Figure** 3.18**a**). We also found that genes with euchromatic H3K4me3 histone modification within 1kb of the TSS have low CpG methylation, and genes with heterochromatic

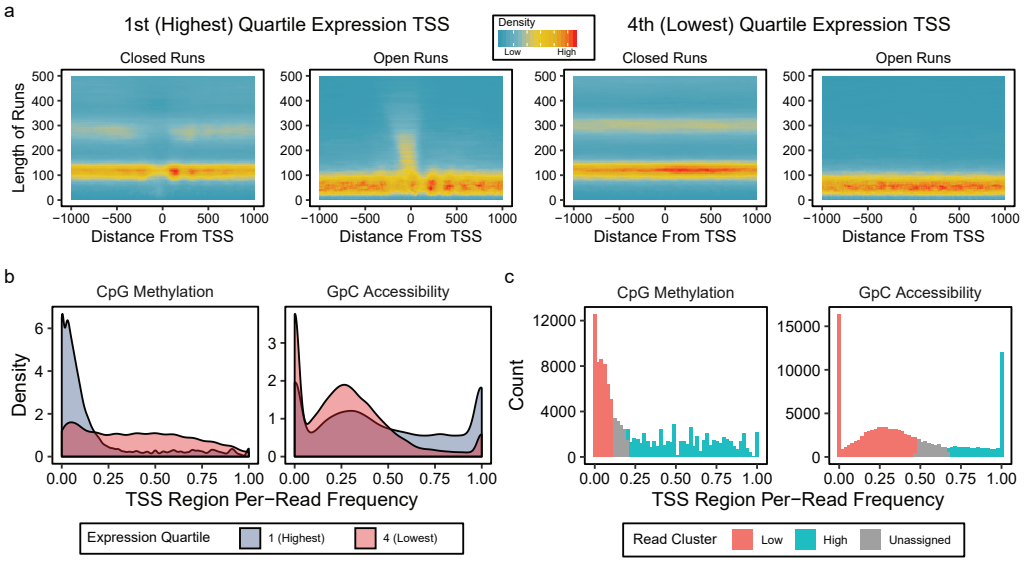


Figure 3.7: Single-read epigenetic assessment on transcription start sites. (a) Heatmaps of lengths of closed and open accessibility runs on individual reads with respect to the distance to transcription start sites, showing the difference in accessibility patterns in highly expressed and lowly expressed gene TSS. **(b)** Distributions of per-read CpG methylation frequency in 1kb region around TSS and GpC accessibility frequency in 200 bp region around TSS, stratified by expression, showing that more reads are demethylated and accessible with an increase in expression. **(c)** We used the windows in (b) to cluster the reads based on methylation around TSS into two groups for each feature, high frequency (blue) and low frequency (red), resulting in four possible combinatorial states for each read at TSS.

H3K27me3 modification mostly have inaccessible reads (**Supplementary Figure 3.18b**). Further, the majority of reads on promoter regions with bivalent histone modifications (both H3K4me3 and H3K27me3) have both low CpG methylation and low accessibility, combining the pattern of CpG methylation with H3K4me3 and that of accessibility with H3K27me3.

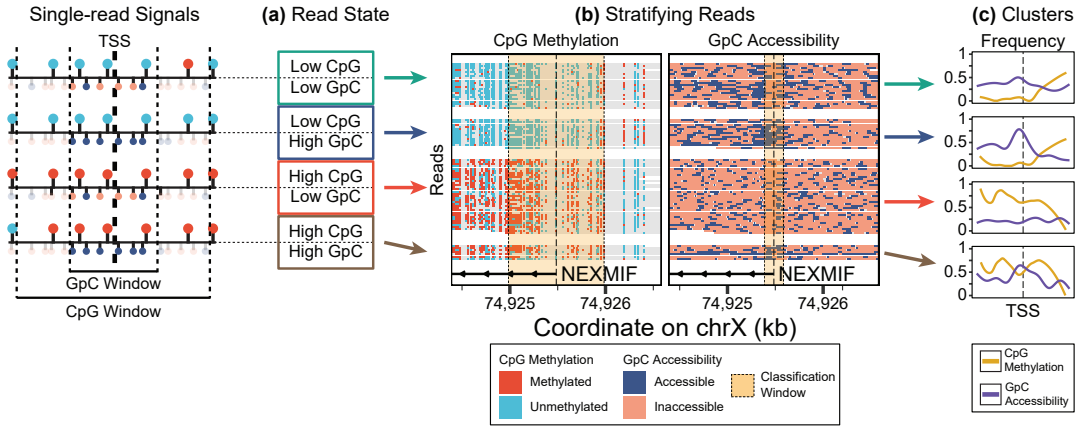


Figure 3.8: Clustering reads based on promoter combinatorial epigenetic state. Reads on TSS can be **a** classified based on the combinatorial epigenetic states near TSS, **b** stratified based on these classifications. **c** Using the long reads, methylation and accessibility frequency profiles can be obtained from each group of combinatorial epigenetic states extending further out from the TSS

3.3.5 Protein binding in association with promoter epigenetic state

We then coupled our ability to probe protein binding and determine combinatorial promoter epigenetic states to identify protein binding events associated with specific promoter states. Using the same subset of genes from the single-read promoter analysis, we examined protein-binding sites within 10kb upstream and downstream of the TSS. We predicted protein binding state on all closed accessibility runs, and selected regions that have multiple overlapping instances of protein binding as candidate sites for protein binding (regions having >= 10 overlapping instances estimated protein-binding states, (see **Methods** 3.5.5). We then performed motif enrichment analysis using Haystack against the JASPAR transcription factor database to determine enrichment values, if any, of these candidate regions in transcription factor

binding sites (**Supplementary Figure 3.19**) (Pinello, Farouni, and Yuan, 2018; Fornes et al., 2020). Several TFs were enriched in the candidate regions, including CTCF, NRF1, and Zinc finger proteins, with the strongest enrichment in CTCF binding sites having a 4x observed/expected ratio. We then stratified the reads based on promoter epigenetic state and calculated the fraction of protein-bound reads in each cluster of combinatorial epigenetic state. In general, groups of reads which had an accessible promoter had a higher fraction of protein-bound reads than inaccessible groups, showing that our protein binding analysis captures protein binding events that are associated with active promoter state (**Figure 3.9**).

For a specific example, we examined PIM2, a gene that facilitates cell survival and proliferation and is highly expressed in GM12878 (1st quartile). PIM2 has a closed run-predicted protein binding site 1.5kb downstream of the TSS present only in the reads with an epigenetically active promoter (**Figure 3.10**). We identified a CTCF binding motif in this region and confirmed it had a peak in existing CTCF ChIP-seq data. This directly links CTCF binding on the same molecule as an accessible promoter 1.5kb away.

We then interrogated individual reads in promoter regions of the breast cancer cells. On promoters of differentially expressed genes in MCF-7/MDA-MB-231 in comparison to MCF-10A, we estimated protein binding and determined single-read promoter epigenetic states. One such gene, ZNF714, is a protein-coding gene for the zinc finger protein 714 and is upregulated in MCF-7 and MDA-MB-231. Two groups of combinatorial epigenetic states are observed in its TSS for all three cell lines: the active (unmethylated and

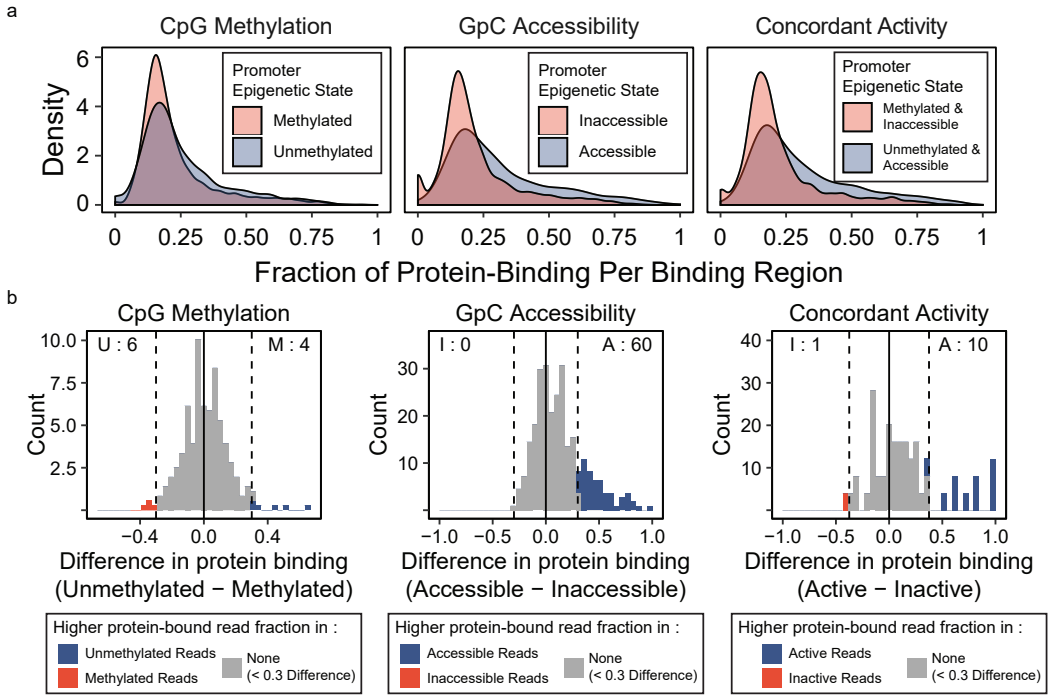


Figure 3.9: Comparisons of predicted TF-binding with respect to promoter epigenetic states. For each of the 1,000 randomly selected genes, protein-binding regions were predicted within 10kb of the TSS. Reads were then split into two groups based on their epigenetic state: (left) CpG methylation, (center) accessibility (right) concordance of both, and the fraction of protein-occupied reads was calculated. The relationship between protein-binding and promoter epigenetic state is assessed by **a** the distribution of fractions of protein-bound reads between read groups, and **b** the difference between the two groups for each region.

accessible) and inactive (methylated and inaccessible) (**Figure** 3.11). The two cancer subtypes have more of the reads in the active state, suggesting that there are more epigenetically active copies of the gene in the cancer subtypes. In addition, the active reads in the cancer subtypes have short closed runs at the same region, suggesting protein binding in these copies, while the active reads in MCF-10A do not. These observations collectively suggest that the up-regulation of ZNF714 occurs in conjunction with increased epigenetic activity and protein binding events near the TSS.

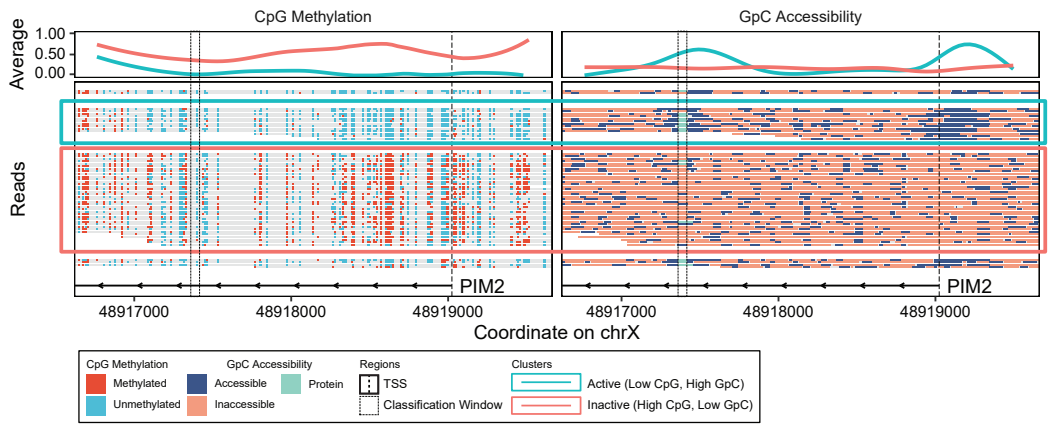


Figure 3.10: Per-read plot of methylation and accessibility on a gene promoter with protein binding. Read-level plots of methylation and accessibility on the promoter of PIM2, showing TF(CTCF) binding 1.5kb downstream of the gene only in concordantly active reads.

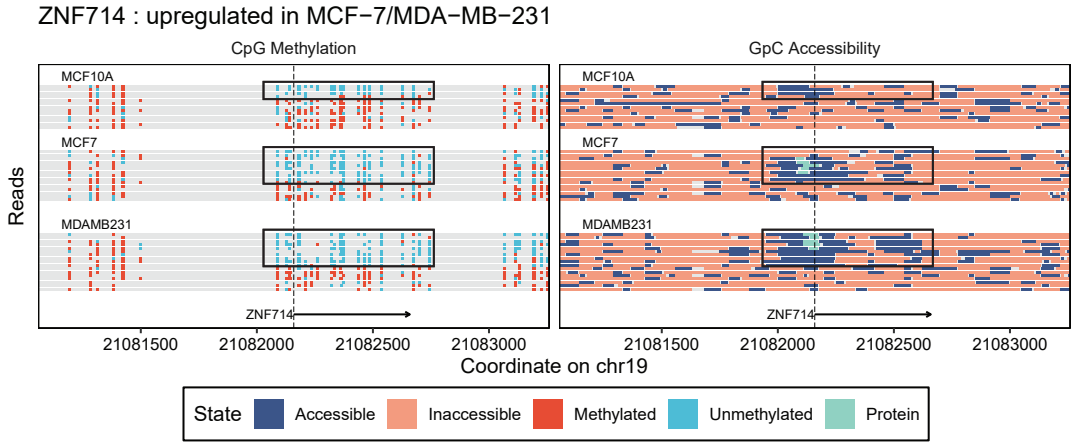


Figure 3.11: Read-level comparative epigenomic analysis of breast cancer model. Read-level methylation and accessibility plots on the TSS of ZNF714 gene, which is upregulated in MCF-7 and MDA-MB-231 in comparison to MCF-10A, showing the differences in the estimated protein binding and combinatorial epigenetic states.

3.3.6 Allele-specific methylation and chromatin accessibility in X chromosome inactivation

Using existing variant data on GM12878 and both parents, we selected heterozygous SNPs and assigned haplotype origin to individual nanoNOMe

reads (Eberle et al., 2016). We were able to confidently determine haplotype assignments on 65% of our sequencing reads across all chromosomes, divided equally to the two haplotypes (paternal: maternal ratios between 0.96 and 1.02), and the phased reads covered 86% of the genome to at least 10x coverage on both alleles (**Figure 3.12**). Having separated the reads based on the parent of origin, we could separate the epigenetic signal within each read into the two alleles as well, therefore generating the first genome-wide allele-specific profiles of DNA methylation and accessibility on a human genome, using a single assay.

We looked into X chromosome inactivation to validate the allele-specific epigenetic profiles. We compared methylation and accessibility near TSSs of autosomal genes, X-chromosome inactivated (XCI) genes, and X-chromosome genes that are known to escape XCI (hereafter referred to as escape genes) via metaplot analysis (**Figure 3.13a**) (Tukiainen et al., 2017). Genes on the active X chromosome (Xa; maternal allele) were concordantly active with demethylated and accessible promoters and those of inactive X chromosome (Xi; paternal allele) were concordantly inactive with methylated and inaccessible promoters, whereas in autosomal genes and escape genes the two alleles had no significant difference in aggregate (**Figure 3.13b**).

3.3.7 Genome-wide allele-specific epigenome analysis

We then found regions that have a significant difference in methylation or accessibility between paternal and maternal alleles, resulting in 9,997 differentially methylated regions (DMRs) and 10,414 differentially accessible regions

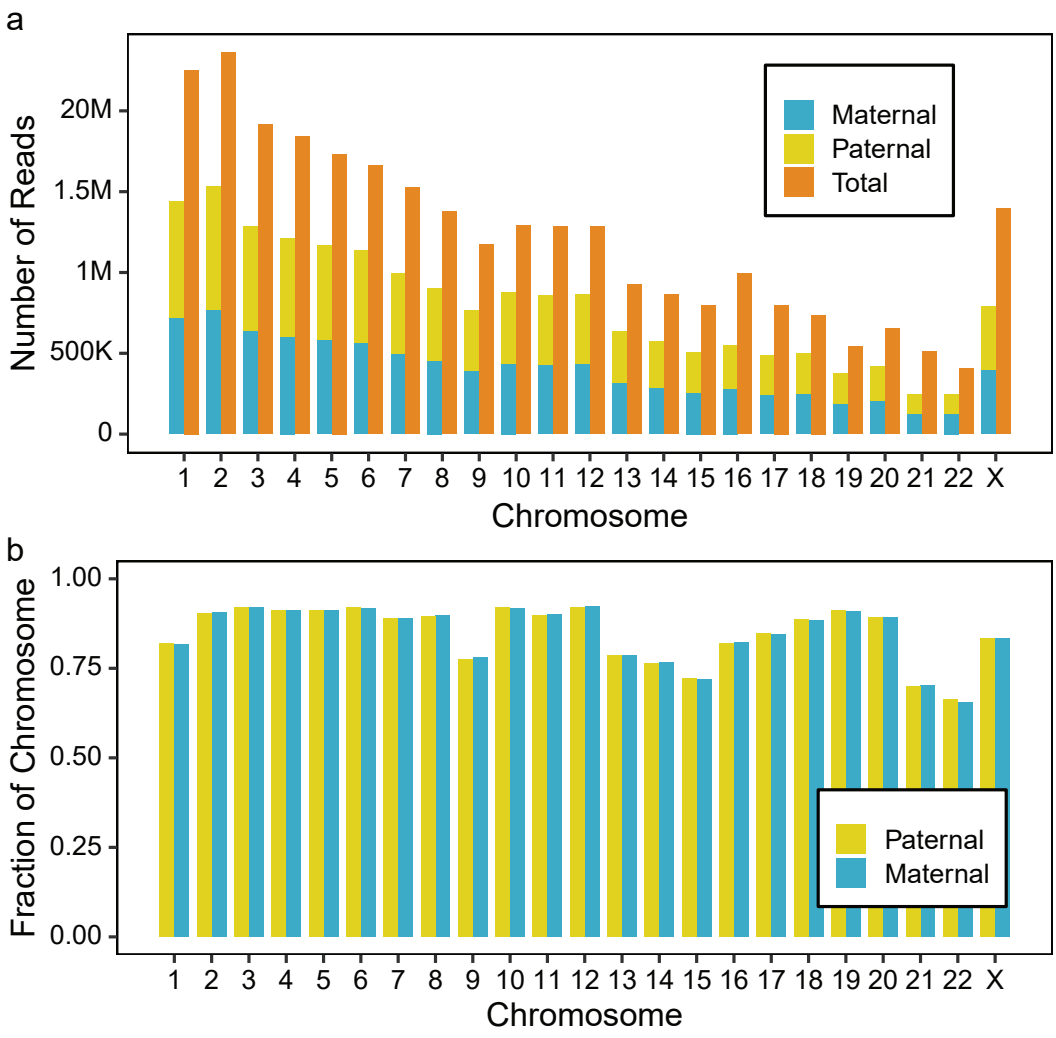


Figure 3.12: Haplotype phasing results on GM12878 nanoNOMe data. **(a)** The number of reads that could be phased into maternal or paternal read based on the presence of heterozygous SNV in the read, showing that 65% of reads could be phased. **(b)** The fractions of the chromosomes that could be phased (the fraction that had > 10x coverage on each allele after phasing) shows on average, 86% of the genome could be phased.

(DARs) across the genome (**Supplementary Figure 3.20a**). While overlaps between DMRs and DARs were not common (629 overlaps, 6%), the overlapping regions showed strong concordance (**Supplementary Figure 3.20b**). In the X chromosome, we observed a disproportionate number of hypermethylated

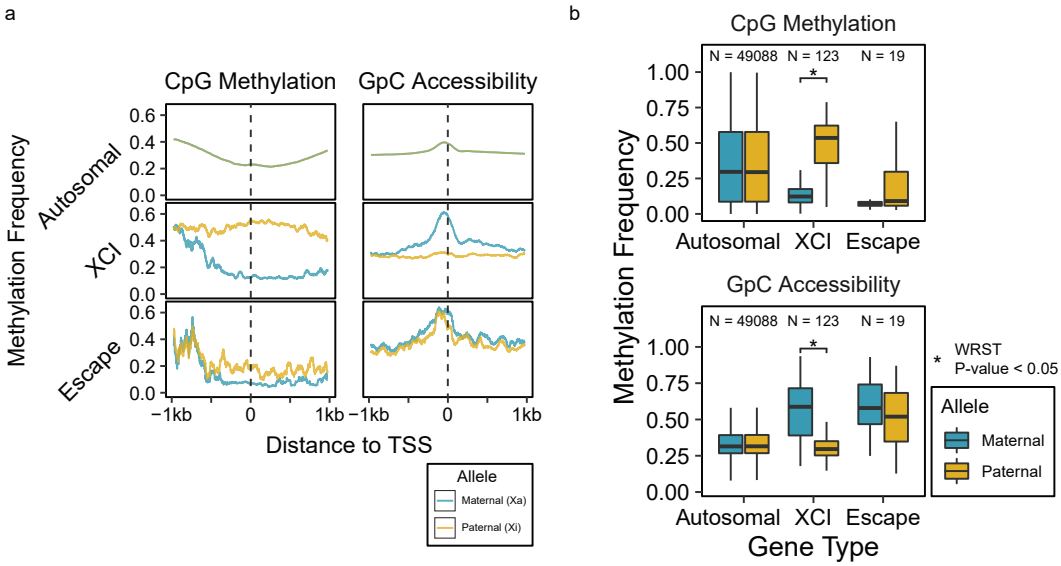


Figure 3.13: Allele-specific epigenetics in X chromosome inactivation. Methylation and accessibility were separated by parent of origin, and **(a)** metaplots of TSS methylation and accessibility were generated for each allele, showing the difference in XCI, and **(b)** boxplots of methylation and accessibility in 500 bp and 100 bp windows, respectively, centered at TSS compared between maternal and paternal alleles.

Xa DMRs (4564 hyper- vs 401 hypo-), agreeing with previous findings that Xa is hypermethylated compared to Xi (Hellman and Chess, 2007). We also found that the majority (N=1050; 80%) of DARs had higher accessibility in Xa, showing that inactivation results in higher overall accessibility of Xa.

To assess the genomic context of DMRs, DARs, and concordant differential regions in XCI, we calculated the enrichment of these regions in various genomic contexts in the X chromosome (**Figure** 3.14). The enrichment of DMRs with higher Xi methylation near TSS (500 bps upstream and downstream) and the high number of DMRs in gene bodies with hypermethylated Xa agreed with previous findings (**Supplementary Figure** 3.21**a**) (Hellman and Chess, 2007; Sharp et al., 2011). However, we found that the high number of DMRs in

gene bodies was due to the larger size of gene bodies, and hypermethylated Xa DMRs were enriched in enhancers. DARs mostly had higher accessibility in Xa, and this pattern was consistent in all assessed genomic contexts. DARs were enriched in CTCF binding sites in addition to promoters, suggesting that the higher accessibility, and consequently increased affinity for CTCF binding, work in concert to prevent XCI in Xa. Concordant regions with both a DAR and DMR were heavily enriched near TSSs and 90% of them indicated higher activity in Xa(307 out of 339). In autosomes, DMRs, DAR, and concordant differential regions all occurred mostly in gene bodies and around TSS, with highest enrichment around TSSs (**Supplementary Figure 3.21b,c**).

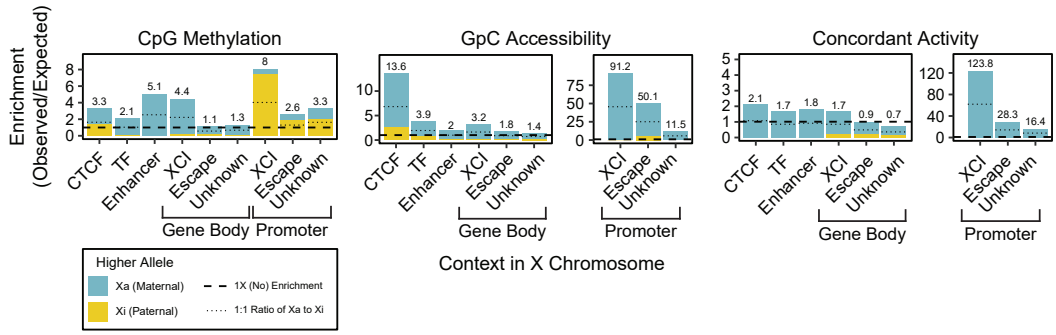


Figure 3.14: Enrichment of allele-specific differential epigenetic regions. Enrichment of (left) DMRs, (center) DARs, and CDRs (right) were calculated at various genomic contexts in the X chromosome, showing the enrichment of allele-specific epigenetic patterns in promoters and regulatory elements.

We then identified genes that had a DMR or a DAR within 500 bp of the TSS. 1,049 genes had a DMR, 868 genes had a DAR, and 245 of these genes had concordant difference in methylation and accessibility near the TSS. 76% (187) of the concordantly differential TSS were in the X chromosome, and all but XIST, a gene known to be specifically active in Xi to promote inactivation of Xi, and RF01880, an exon of the XIST gene, in chrX indicated activity

in Xa (maternal allele). Out of the 56 autosomal genes, 8 were previously identified imprinted genes (Jirtle, 1999; Morison, Ramsay, and Spencer, 2005). We plotted ZNF597, one of the 8 known imprinted genes, as an example; it had a hypermethylated and less accessible promoter in the maternal copy, indicating that it is active in the paternal allele (**Figure 3.15**). In addition to the TSS epigenetic states, we observed that the gene body exhibited the opposite pattern of methylation, with the active paternal copy being fully methylated.

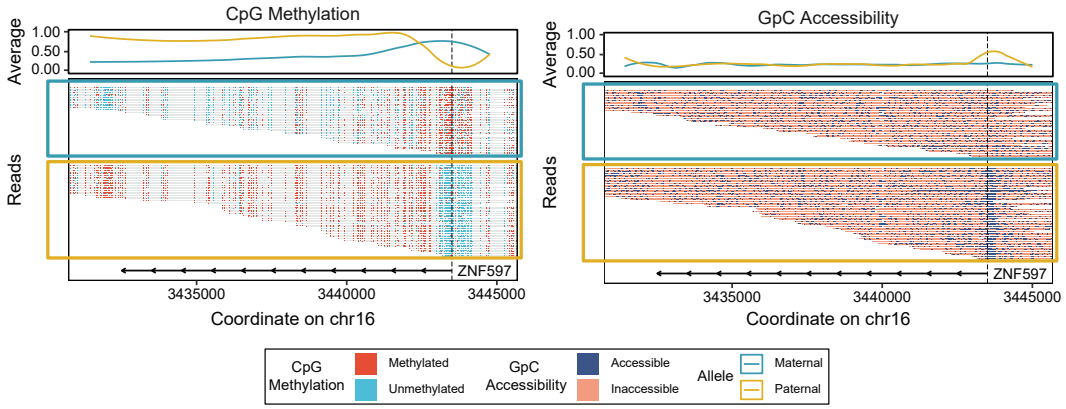


Figure 3.15: Allele-specific per-read methylation and accessibility of an ZNF597. Read-level methylation and accessibility plots of ZNF597, an imprinted gene with allele-specific epigenetic patterns at the promoter of the gene.

3.3.8 Allele-specific epigenomics in heterozygous structural variations

Our long nanopore reads also allow detection of structural variants, large insertions, deletions, or transpositions hard to detect with conventional short-read sequencing. We characterized epigenetic consequences of these SVs by comparing epigenetic signals in heterozygous SVs, focusing on large deletions and insertions, which were the most commonly occurring SV types. After

filtering for SVs that strongly suggest heterozygous SVs (filtering method outlined in (**Methods** 3.5.6), we identified 1,195 deletions and 1,167 insertions, and compared methylation and accessibility near SV breakpoints between the variant and reference alleles (**Figure 3.16, Supplementary Figure** 3.22). We found that while the majority of the SVs (80% of deletions and 82% of insertions) do not have a difference in methylation between the alleles, in those that did have a difference, the variant allele tends to be hypomethylated in deletions (173 hypo vs. 65 hyper-) and hypermethylated in insertions (84 hypo vs. 131 hyper). This suggests a relationship between structural variation and epigenetic state, and that nanoNOMe has the capability to detect these changes.

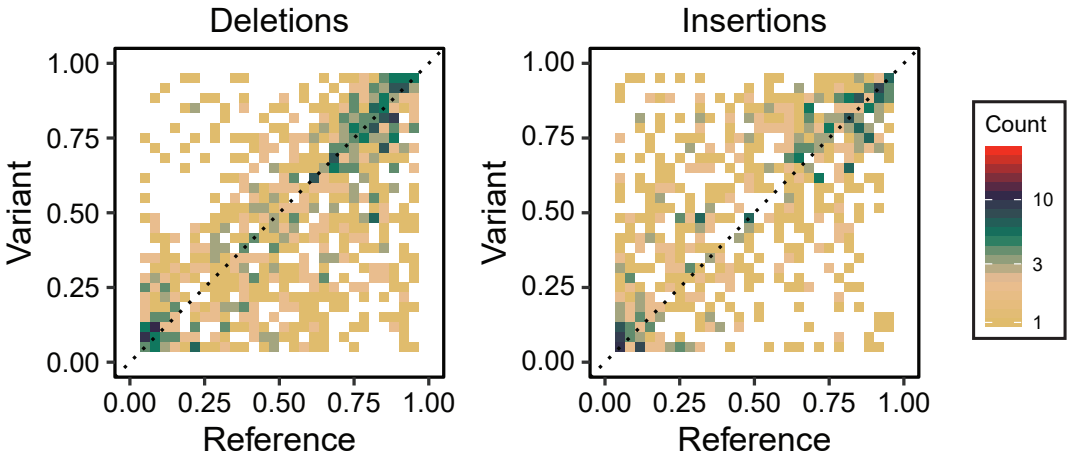


Figure 3.16: CpG methylation in heterozygous structural variations. Pairwise comparison of reference and variant allele CpG methylation in 1 kb region around breakpoints of heterozygous structural variations.

3.4 Discussion

We have utilized the long reads obtained by nanoNOMe to explore a number of aspects in the epigenome. Accessibility signals from long reads span multiple protein binding sites, generating footprints with lengths which allow us to infer the type of protein occupying the region. Using this approach coupled to known CTCF binding motifs, we have examined the relationship between CTCF binding and epigenetic patterns of nearby regions. Further, we have combined the ability to predict protein binding with combinatorial promoter epigenetic states to show protein binding events that occur in cis with specific promoter epigenetic states. We can use these tools to identify differential epigenetic states and protein binding events between different breast cancer cell lines, providing a new window on cancer gene regulation. The ability to observe long-range interactions of epigenetic features will be useful in studying epigenomes in heterogeneous populations and in aneuploid genomes (Matzke, Mittelsten Scheid, and Matzke, 1999). Because of the greater chance for long reads to encounter a heterozygous SNV, we can phase our nanopore reads, generating fully phased methylation and accessibility profiles of a human genome. We explored the phased X chromosome to understand new features of the epigenetic profile of X-inactivation. Curiously, the inactive X chromosome showed lower methylation than the active X chromosome outside of the TSS regions (which were more methylated in Xi), especially in enhancers. We directly demonstrated that the allele-specific data is useful in observing parent-of-origin epigenetic features, such as X chromosome inactivation and escape from inactivation, allele-specific activity

of imprinted genes, and epigenetic differences near heterozygous structural variations. We can also use such tools to explore how imprinting is initiated and controlled, by examining the phased epigenome in different tissues and different developmental stages. Lastly, we can phase heterozygous SVs with our long reads, and compare the epigenome of alleles with and without the SV. The ability to measure the phased epigenome will be of high utility for exploring allele-specific epigenetic states, a recognized feature of human cancers (Tischoff et al., 2005; Avin et al., 2019).

3.5 Methods

3.5.1 Calculating and piling up co-occurrence of accessibility and inaccessibility

To observe patterns of DNA methylation and accessibility in the presence of biological heterogeneity and technical variability, co-occurrence of methylated/unmethylated cytosine is calculated across reads that map to the genomic region of interest. Co-occurrence, c, is defined by the same event, M (methylated or unmethylated) , occurring on two separate binned locations, i and j, along a given read :

$$c_{ij} = \begin{cases} 1, & \text{if } M_i \geq M_j \\ 0, & \text{otherwise} \end{cases}$$

After calculating the co-occurrence for each pair of coordinates for each read, the counts are piled up to determine the frequency of co-occurrence as a measure of how often reads have the same events occurring between the

positions i and j. The resulting matrix of co-occurrence pileup is normalized by the maximum count, and plotted as a 2-dimensional heatmap to visualize the patterns (see accessions for code availability).

3.5.2 Estimating single-molecule accessibility calls using a smoothing estimator

To remove isolated erroneous calls of accessibility on individual reads, we applied a fixed-bandwidth Gaussian kernel smoothing on the log-likelihood ratios (LLRs). First, LLR were capped at the calling thresholds (-1,1), forcing all LLRs with absolute value greater than 1 to be -1 or 1, to prevent bias from LLRs with large magnitudes. The adjusted LLRs were smoothed using a Gaussian kernel with fixed bandwidths. Smoothed LLRs were called as accessible (methylated) if the LLR > 0.4 and inaccessible (unmethylated) if the LLR < 0.4.

3.5.3 Predicting regulatory protein binding from closed runs

To discriminate CTCF binding events from nucleosome binding events on individual reads, we used lengths of closed runs on centers of CTCF binding motifs. The lengths were clustered on Gaussian finite mixture models using Expectation-Maximization algorithm implemented by R package Mclust version 5.4.5 (Scrucca et al., 2016). The optimal clustering parameters were determined based on maximum integrated complete-data likelihood (ICL), and the cluster that had the smallest mean length (54 bps) was chosen as CTCF-binding signal and the other clusters as units of nucleosome-binding signals. This model was applied to classify all closed runs within 25 bps of

CTCF binding sites as CTCF-bound or nucleosome-bound and reads that contained CTCF-bound closed runs were considered to be CTCF-bound reads. To predict protein binding events outside of CTCF binding sites, we used the model on all closed runs to categorize them to one of the clusters, using the runs that were assigned into the smallest mean length cluster as candidates for protein binding. We selected regions that contained at least ten candidates as the predicted regions of protein binding events.

3.5.4 Predicting combinatorial promoter epigenetic states on individual reads

To predict combinatorial epigenetic states of individual reads on TSS, we used methylation and accessibility in a window around the TSS. On each read that spans a TSS, we calculated average CpG methylation over the 1kb region around the TSS and GpC methylation over 200 bps around the TSS. The two epigenetic signals were separately clustered into two clusters of high and low average signals using the Expectation-Maximization algorithm on Gaussian finite mixture models as used above to generate probabilistic models of the epigenetics states. On individual reads, CpG methylation and GpC accessibility were separately clustered using the resulting models, and the combinatorial epigenetic state of reads was determined based on the combination of the cluster assignments.

3.5.5 Interactions of promoter epigenetic states and protein binding

Read-level protein-binding estimation and promoter epigenetic state estimation were coupled by first estimating regions of protein binding within 10kb of TSSs of a subset of genes. Sites that have 10 or more reads with short closed runs in a window less than 80 bp were selected as estimated protein-binding regions. We then separated the reads based on the epigenetic state of nearby gene promoter(s), and separately assessed the reads that suggest a protein-binding event at the protein-binding region in each group, resulting in protein-bound reads specific to each promoter epigenetic state.

3.5.6 Haplotype Assignment and Allele-Specific Methylation Analysis

We obtained genotype information for GM12878 from existing phased Illumina platinum genome data generated by deep sequencing of the cell donors' familial trio (Eberle et al., 2016). Bcftools version 1.9 was used to filter for only variants that are heterozygous in GM12878 (Li, 2011). The heterozygous GM12878 SNVs were used to identify reads with allele-informative variants and assign the parent of origin for each read using WhatsHap version 0.18 (Patterson et al., 2015). Methylation and accessibility calls on each read were separated based on the haplotype assignments to generate allele-specific profiles of methylation and accessibility. To identify accurate heterozygous SVs, we called SVs on the two alleles separately using Sniffles version 1.0.11 and SURVIVOR version 1.0.7 using default parameters (Sedlazeck et al., 2018).

From the resulting merged vcf, we selected heterozygous SVs by selecting the SVs that have less than 2 non-variant and more than 20 variant reads on only one of the alleles. To remove SVs that are short in length or affected by incorrect alignments, we removed SVs that are shorter than 200 bps and have more than 100 read alignments in one allele.

3.5.7 Bresat cancer cell line analysis

RNA-seq counts of the three cell lines were downloaded from GEO accession GSE75168 and analyzed using the bioconductor packager DESeq2 version 1.24.0 (Messier et al., 2016; Love, Huber, and Anders, 2014). Using default parameters, differential expression analysis was performed based on the negative binomial distribution, comparing MCF-7 and MDA-MB-231 to MCF-10A. Genes were considered to be significantly differentially expressed when the Bonferroni-Hochberg corrected p-values were less than 0.01.

Of this differentially expressed set, we filtered for genes with 5 or more differences between normal and cancer lines in the number of reads indicating epigenetically concordant active promoters. Then the protein binding states were compared on predicted protein-binding regions within 10kb of the TSS of these genes, and the fraction of reads that have protein binding were calculated for each sample. Those regions that had a difference of protein binding fraction >= 0.4 were selected as genes with differences in the promoter epigenetic states and protein binding.

3.6 Acknowledgments

Funding: This study was supported by National Human Genome Research Institute (NHGRI project 5R01HG009190)

Competing Interests: WT has two patents (8,748,091 and 8,394,584) licensed to Oxford Nanopore. JTS received research funding from ONT. IL, TG, NS, JTS, FJS and WT have received travel funds to speak at meetings from ONT.

Data Accessions: NanoNOMe data of GM12878, MCF-10A, MCF-7, and MDA-MB-231 are available at NCBI Bioproject ID PRJNA510783 (http://www.ncbi.nlm.nih.gov/bioproject/510783). Source code is available at https://github.com/timplab/nanoNOMe.

3.7 Supplementary Material

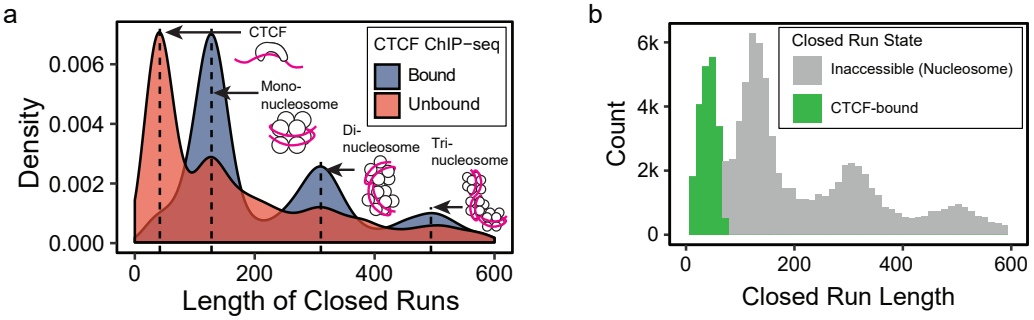


Figure 3.17: Single-molecule closed run lengths at CTCF binding sites. (a) Density distributions of closed runs at the CTCF binding sites, showing that sites without CTCF binding have long closed runs suggesting nucleosome binding while those with CTCF binding have short closed runs suggesting CTCF binding, and **(b)** using the length of runs at CTCF binding sites to discriminate inaccessibility due to regulatory protein binding from nucleosome binding.

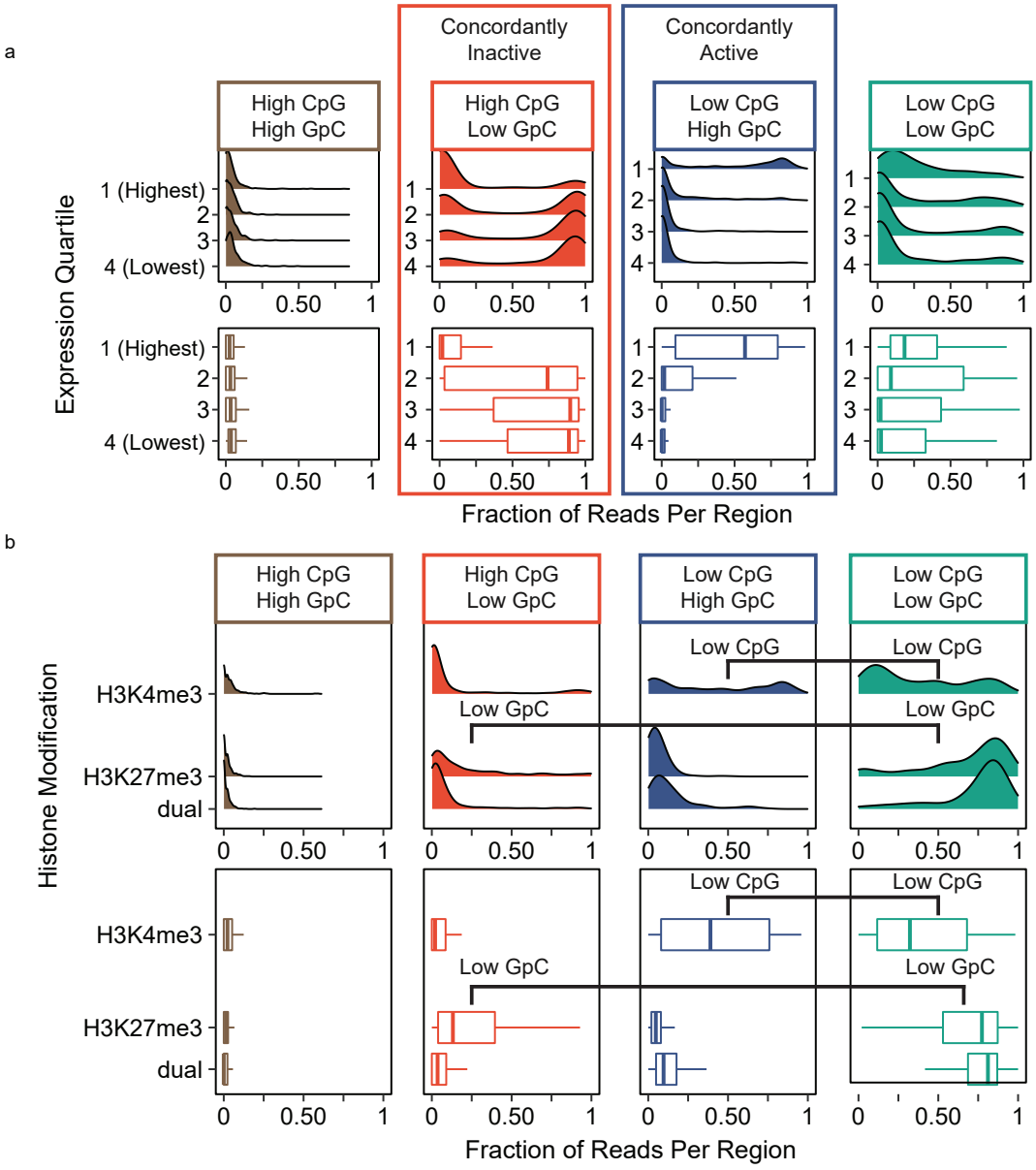
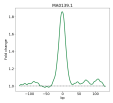
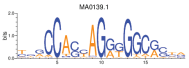
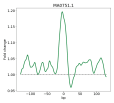
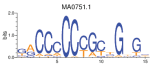
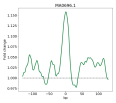
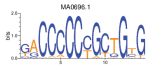
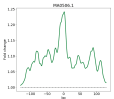
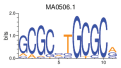
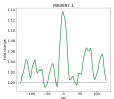
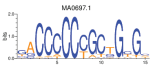
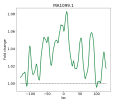
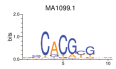
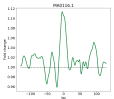
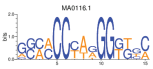
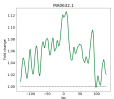
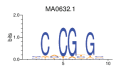
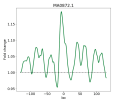
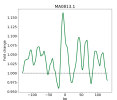
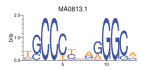


Figure 3.18: Assessment of read-level combinatorial epigenetic states of TSS. Fractions of read-level combinatorial epigenetic states at TSS were calculated for each TSS in a subset of 1,000 genes per group and compared **(a)** by expression quartiles, showing that with increasing expression more genes have higher fraction of combinatorially active reads and less fraction of inactive reads, and **(b)** by promoter histone modification, showing that reads at euchromatic H3K4me3 genes are demethylated and reads at heterochromatic H3K27me3 genes are inaccessible.

Haystack Motif Enrichment Analysis Tool

Motif ID	Motif Name	Presence in Target	Presence in BG	Ratio	p-value	q-value	Central Enrichment	Motif Profile	Logo
MA0139.1	CTCF	16.05%	3.98%	3.42	2.08e-112	6.87e-111	3.82		
MA0751.1	ZIC4	7.97%	4.76%	1.56	2.13e-12	1.41e-11	1.81		
MA0696.1	ZIC1	7.14%	4.07%	1.61	6.80e-13	5.61e-12	1.76		
MA0506.1	NRF1	7.70%	3.61%	1.89	1.02e-21	1.69e-20	1.76		
MA0697.1	ZIC3	6.93%	4.76%	1.38	9.33e-07	3.08e-06	1.66		
MA1099.1	Hes1	3.55%	1.59%	1.75	1.99e-11	1.10e-10	1.66		
MA0116.1	Znf423	3.14%	2.16%	1.31	1.18e-03	2.05e-03	1.64		
MA0632.1	Tcfl5	3.02%	1.06%	1.95	3.17e-14	3.48e-13	1.59		
MA0872.1	TFAP2A(var.3)	5.29%	3.95%	1.27	6.94e-04	1.28e-03	1.58		
MA0813.1	TFAP2B(var.3)	5.07%	3.84%	1.25	1.56e-03	2.57e-03	1.55		

Haystack was built by Luca Pinello

Figure 3.19: Results of Haystack Bio motif enrichment on candidate protein binding sites. Candidate protein binding sites were selected from closed accessibility runs. Haystack Bio motif enrichment was ran on these regions to detect enrichment of these regions in transcription factor binding sites from JASPAR database.

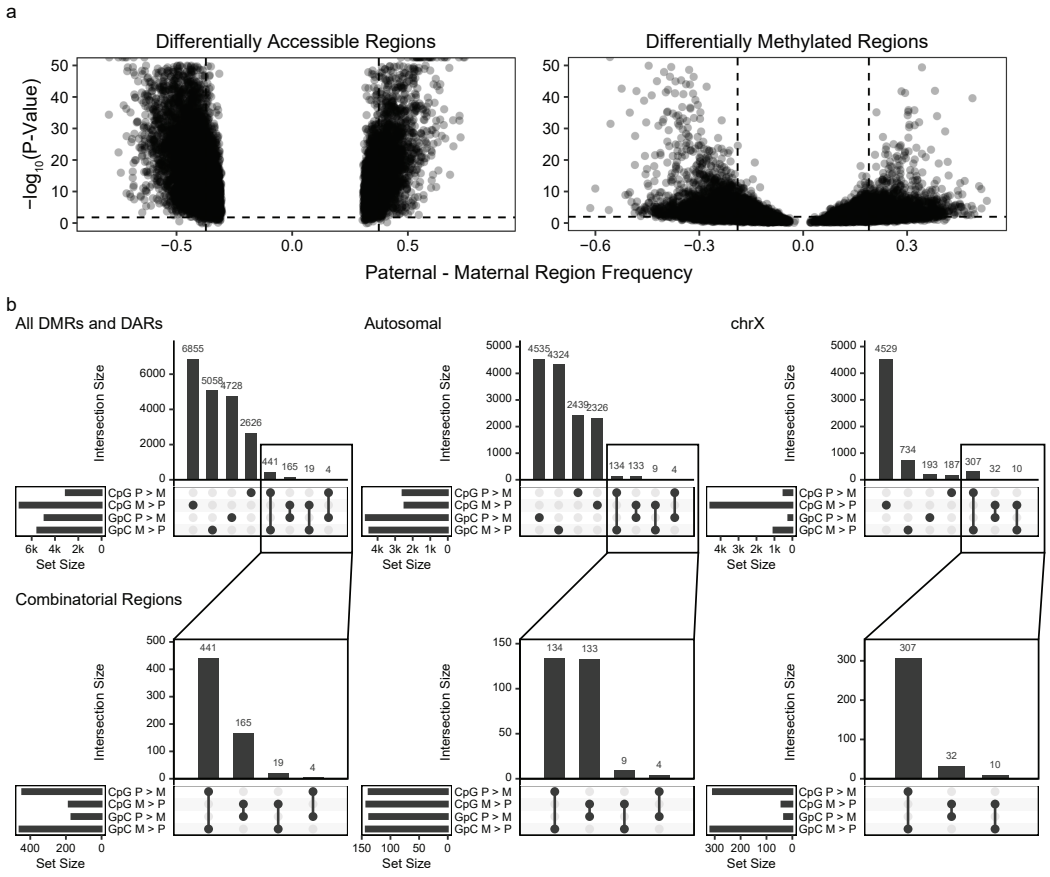


Figure 3.20: Differentially methylated and differentially accessible regions between alleles in GM12878. Methylation was compared between the two alleles across the genome to find regions of significant difference and were tested using one-sided FisherâĂŹs exact test, and accessibility peaks were compared by 1) finding peaks of accessibility on each allele separately, 2) selecting peaks that occur exclusively in one allele, 3) and comparing the accessibility frequency between the two alleles in these candidate regions. The detected DMRs and DARs are **(a)** shown as volcano plots, with dashed lines representing thresholds for considering the region as DMR/DAR. **(b)** Directions of DMRs/DARs as well as their overlaps were observed using upset plots, across the genome (left), as well as separated by autosomes (middle) and X chromosome (right).

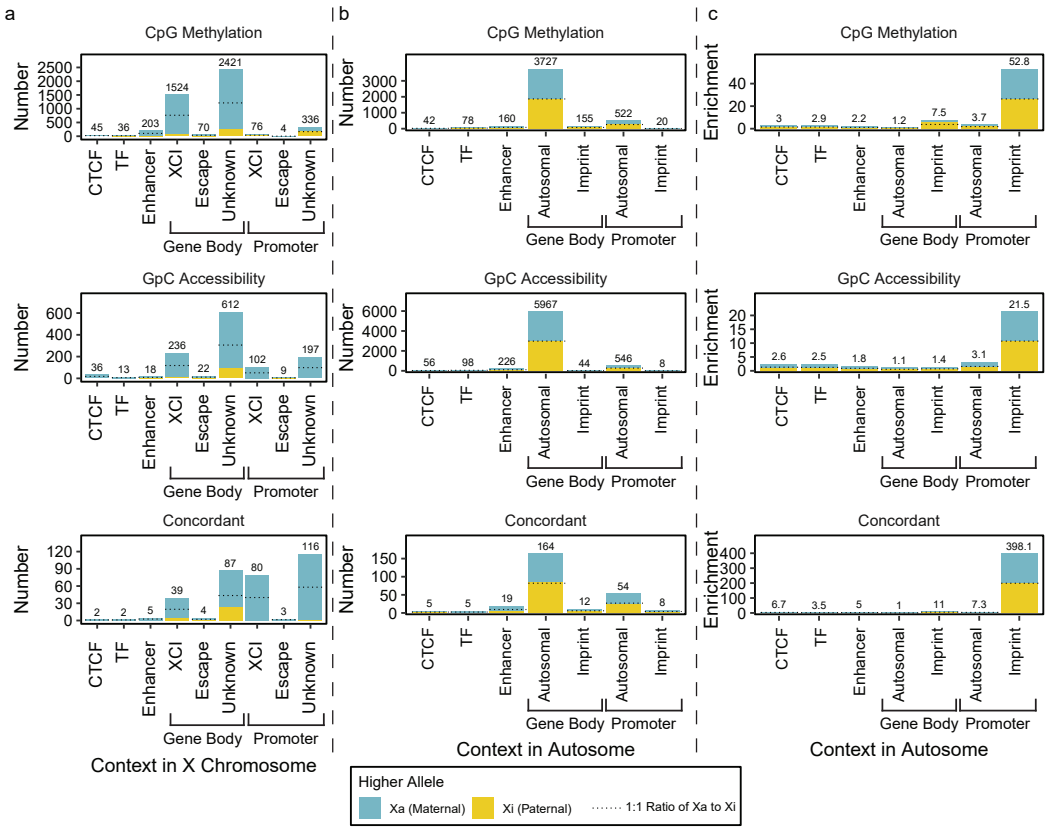


Figure 3.21: Genome-context enrichment of allele-specific differential epigenetic regions. **(a)** Numbers of differential regions in each genomic context in X chromosome, **(b)** Numbers of differential regions in genomic contexts in autosomes, and **(c)** enrichment of differential regions in autosomes

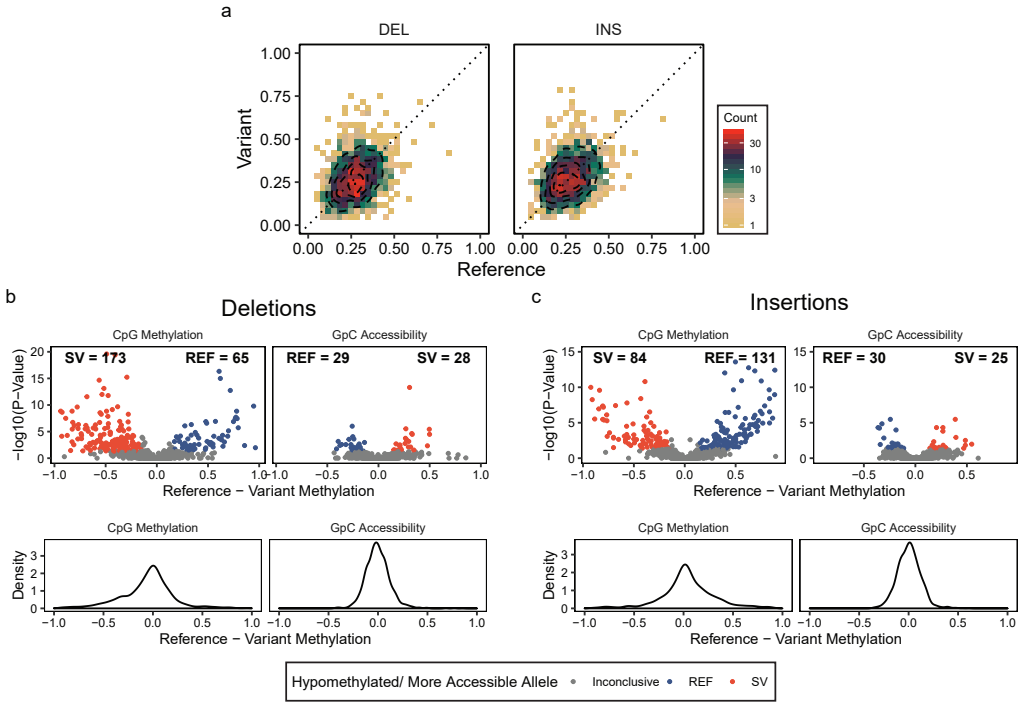


Figure 3.22: Allele-specific epigenetic comparison of heterozygous structural variations. (a) Pair-wise comparison of GpC methylation around breakpoints of heterozygous SVs, with the allele with the SV on the y-axis and the allele without the SV on the x-axis. The difference of CpG methylation and GpC methylation in variant alleles of SVs in comparison to the reference alleles, with one-sided FisherâĂŹs exact tests and presented as volcano plots for **(b)** heterozygous deletions and **(c)** heterozygous insertions.

References

Buenrostro, Jason D, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf (2015). "Single-cell chromatin accessibility reveals principles of regulatory variation". en. In: *Nature* 523.7561, pp. 486–490.

Lai, Binbin, Weiwu Gao, Kairong Cui, Wanli Xie, Qingsong Tang, Wenfei Jin, Gangqing Hu, Bing Ni, and Keji Zhao (2018). "Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing". en. In: *Nature* 562.7726, pp. 281–285.

Guo, Hongshan, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang (2013). "Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing". en. In: *Genome Res.* 23.12, pp. 2126–2135.

Smallwood, Sébastien A, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey (2014). "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity". en. In: *Nat. Methods* 11.8, pp. 817–820.

Clark, Stephen J, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, Oliver Stegle, and Wolf Reik (2018). "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells". en. In: *Nat. Commun.* 9.1, p. 781.

Lai, William K M and B Franklin Pugh (2017). "Understanding nucleosome dynamics and their links to gene expression and DNA replication". en. In: *Nat. Rev. Mol. Cell Biol.* 18.9, pp. 548–562.

Satpathy, Ansuman T, Jeffrey M Granja, Kathryn E Yost, Yanyan Qi, Francesca Meschi, Geoffrey P McDermott, Brett N Olsen, Maxwell R Mumbach, Sarah E Pierce, M Ryan Corces, Preyas Shah, Jason C Bell, Darisha Jhutty, Corey M Nemec, Jean Wang, Li Wang, Yifeng Yin, Paul G Giresi, Anne Lynn S Chang, Grace X Y Zheng, William J Greenleaf, and Howard Y Chang (2019).

"Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion". en. In: *Nat. Biotechnol.* 37.8, pp. 925–936.

Han, Li, Dong-Hoon Lee, and Piroska E Szabó (2008). "CTCF is the master organizer of domain-wide allele-specific chromatin at the H19/Igf2 imprinted region". en. In: *Mol. Cell. Biol.* 28.3, pp. 1124–1135.

Fournier, Cécile, Yuji Goto, Esteban Ballestar, Katia Delaval, Ann M Hever, Manel Esteller, and Robert Feil (2002). "Allele-specific histone lysine methylation marks regulatory regions at imprinted mouse genes". en. In: *EMBO J.* 21.23, pp. 6560–6570.

Singer-Sam, Judith and Arthur D Riggs (1993). "X chromosome inactivation and DNA methylation". In: *DNA Methylation: Molecular Biology and Biological Significance*. Ed. by Jean-Pierre Jost and Hans-Peter Saluz. Basel: Birkhäuser Basel, pp. 358–384.

Lengauer, C, K W Kinzler, and B Vogelstein (1998). "Genetic instabilities in human cancers". en. In: *Nature* 396.6712, pp. 643–649.

1000 Genomes Project Consortium, Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean (2010). "A map of human genome variation from population-scale sequencing". en. In: *Nature* 467.7319, pp. 1061–1073.

Rozowsky, Joel, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, Nitin Bhardwaj, Mark Rubin, Michael Snyder, and Mark Gerstein (2011). "AlleleSeq: analysis of allele-specific expression and binding in a network framework". en. In: *Mol. Syst. Biol.* 7, p. 522.

Bansal, Vikas and Vineet Bafna (2008). "HapCUT: an efficient and accurate algorithm for the haplotype assembly problem". en. In: *Bioinformatics* 24.16, pp. i153–9.

Ziebarth, Jesse D, Anindya Bhattacharya, and Yan Cui (2013). "CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization". en. In: *Nucleic Acids Res.* 41.Database issue, pp. D188–94.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414, pp. 57–74.

Hesselberth, Jay R, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, Shane Neph, Michael S Kuehn, William S Noble, Stanley Fields, and John A Stamatoyannopoulos (2009). "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting". en. In: *Nat. Methods* 6.4, pp. 283–289.

Luscombe, N M, S E Austin, H M Berman, and J M Thornton (2000). "An overview of the structures of protein-DNA complexes". en. In: *Genome Biol.* 1.1, REVIEWS001.

Boyle, Alan P, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford (2008). "High-resolution mapping and characterization of open chromatin across the genome". en. In: *Cell* 132.2, pp. 311–322.

Pinello, Luca, Rick Farouni, and Guo-Cheng Yuan (2018). "Haystack: systematic analysis of the variation of epigenetic states and cell-type specific regulatory elements". en. In: *Bioinformatics* 34.11, pp. 1930–1933.

Fornes, Oriol, Jaime A Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A Richmond, Bhavi P Modi, Solenne Correard, Marius Gheorghe, Damir Baranašić, Walter Santana-Garcia, Ge Tan, Jeanne Chèneby, Benoit Ballester, François Parcy, Albin Sandelin, Boris Lenhard, Wyeth W Wasserman, and Anthony Mathelier (2020). "JASPAR 2020: update of the open-access database of transcription factor binding profiles". en. In: *Nucleic Acids Res.* 48.D1, pp. D87–D92.

Eberle, Michael A, Epameinondas Fritzilas, Peter Krusche, Morten Kallberg, Benjamin L Moore, Mitchell A Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J Humphray, Aaron L Halpern, Semyon Kruglyak, Elliott H Margulies, Gil McVean, and David R Bentley (2016). "A reference dataset of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". en.

Tukiainen, Taru, Alexandra-Chloé Villani, Angela Yen, Manuel A Rivas, Jamie L Marshall, Rahul Satija, Matt Aguirre, Laura Gauthier, Mark Fleharty, Andrew Kirby, Beryl B Cummings, Stephane E Castel, Konrad J Karczewski, François Aguet, Andrea Byrnes, GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Tuuli Lappalainen, Aviv Regev, Kristin G Ardlie, Nir Hacohen, and Daniel

G MacArthur (2017). “Landscape of X chromosome inactivation across human tissues”. en. In: *Nature* 550.7675, pp. 244–248.

Hellman, Asaf and Andrew Chess (2007). “Gene body-specific methylation on the active X chromosome”. en. In: *Science* 315.5815, pp. 1141–1143.

Sharp, Andrew J, Elisavet Stathaki, Eugenia Migliavacca, Manisha Brahmachary, Stephen B Montgomery, Yann Dupre, and Stylianos E Antonarakis (2011). “DNA methylation profiles of human active and inactive X chromosomes”. en. In: *Genome Res.* 21.10, pp. 1592–1600.

Jirtle, R L (1999). “Genomic imprinting and cancer”. en. In: *Exp. Cell Res.* 248.1, pp. 18–24.

Morison, Ian M, Joshua P Ramsay, and Hamish G Spencer (2005). “A census of mammalian imprinting”. en. In: *Trends Genet.* 21.8, pp. 457–465.

Matzke, M A, O Mittelsten Scheid, and A J Matzke (1999). “Rapid structural and epigenetic changes in polyploid and aneuploid genomes”. en. In: *Bioessays* 21.9, pp. 761–767.

Tischoff, Iris, Annett Markwarth, Helmut Witzigmann, Dirk Uhlmann, Johann Hauss, Alireza Mirmohammadsadegh, Christian Wittekind, Ulrich R Hengge, and Andrea Tannapfel (2005). “Allele loss and epigenetic inactivation of 3p21.3 in malignant liver tumors”. en. In: *Int. J. Cancer* 115.5, pp. 684–689.

Avin, Brittany A, Yongchun Wang, Timothy Gilpatrick, Rachael E Workman, Isac Lee, Winston Timp, Christopher B Umbricht, and Martha A Zeiger (2019). “Characterization of human telomerase reverse transcriptase promoter methylation and transcription factor binding in differentiated thyroid cancer cell lines”. en. In: *Genes Chromosomes Cancer* 58.8, pp. 530–540.

Scrucca, Luca, Michael Fop, T Brendan Murphy, and Adrian E Raftery (2016). “mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models”. en. In: *R J.* 8.1, pp. 289–317.

Li, Heng (2011). “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data”. en. In: *Bioinformatics* 27.21, pp. 2987–2993.

Patterson, Murray, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W Klau, and Alexander Schönhuth (2015). “WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads”. en. In: *J. Comput. Biol.* 22.6, pp. 498–509.

Sedlazeck, Fritz J, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz (2018). "Accurate detection of complex structural variations using single-molecule sequencing". en. In: *Nat. Methods* 15.6, pp. 461–468.

Messier, Terri L, Jonathan A R Gordon, Joseph R Boyd, Coralee E Tye, Gillian Browne, Janet L Stein, Jane B Lian, and Gary S Stein (2016). "Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes". en. In: *Oncotarget* 7.5, p. 5094.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12, p. 550.

Chapter 4

Targeted sequencing on nanopore sequencing platform

Isac Lee, Rachael Workman, Winston Timp, Josh Zhiyong Wang. Use of Agilent SureSelect to perform targeted long-read nanopore sequencing. Agilent Application Note (2017) doi:10.17504/protocols.io.zxyf7pw

4.1 Abstract

Targeted enrichment of DNA in genomic regions of interest is a cost-effective method to perform deep sequencing on multiple samples. Targeted sequencing is widely used in Next Generation Sequencing (NGS), e.g. exome sequencing. Here I implemented of targeted nanopore sequencing and discuss the challenges and advantages. By adapting solution-phase hybridization capture to nanopore sequencing, we can achieve > 300-fold enrichment in DNA sequences, which allows deeper analysis of genomic variants and can be useful in large-scale studies.

4.2 Introduction

The explosive advances in DNA sequencing technologies has made it a universal tool for biology with applications in a wide range of studies. However, regardless of how cheap sequencing is, a more cost-effective methodology for sequencing the DNA to a great depth is always favorable, especially in organisms with large genomes such as humans. One of the ways to make an efficient use of the sequencing is to selectively sequence parts of the genome that is important in the biological question. In fact, we have not yet uncovered the function of the vast majority of the genome, so much of the data in whole genome sequencing is seemingly wasted (Salzberg, 2019). Several approaches for targeted enrichment have been established, including PCR, tagmentation (fragmentation by transposase), and solution-phase hybridization capture (Kozarewa et al., 2015).

The most obvious benefit of targeted sequencing is the reduction in cost. Because large numbers of samples can be sequenced with low amount of input using targeted sequencing, it has been used in large biobank consortia and clinical studies. Multiple biobanks have sequenced whole exomes of over 1,000 individuals using exome sequencing, associating exon mutations with disease phenotypes (GTEx Consortium et al., 2017; Van Hout et al., 2019). Panels for specifically targeting cancer-associated regions in the genome have been developed and applied to clinical studies, allowing comprehensive mutation studies of cancer driver mutations across large number of patients as well as prevention and early detection of cancers (Nikiforova et al., 2013; Muller et al., 2016; Lee et al., 2015). In addition, targeted sequencing yields high

depth of sequencing, allowing more accurate measurements of mutations in the regions of interest.

One of the most popular target enrichment methods available is from Agilent Technologies: the SureSelectXT solution-phase hybridization-capture system (**Figure** 4.1). In addition to predetermined panels of capture probes, the user can utilize custom capture libraries with 120nt biotinylated RNA baits to enrich genomic regions ranging from less than 50 kb to over 100 Mb for deep sequencing of specific genomic regions. With probe designs generated by their design algorithms, which considers complicated factors such as sequence complexity and GC content, it is possible to perform DNA enrichment and sequencing in a highly efficient, cost-effective manner. Here we present our adaptation of the Agilent Sureselect system on nanopore sequencing to sequence tumor suppressor genes *CDKN2A* and *SMAD4*. We demostrate its ability to deeply sequence regions of interest and use it to detect structural variations and phase single nucleotide variations.

4.3 Results

4.3.1 Solution-phase hybridization capture nanopore sequencing

We have applied the Agilent SureSelectXT protocol to nanopore long-read sequencing. Note that the vast majority of conventional DNA sequencing library preparation protocols are geared toward creating short, 200 - 300 bp DNA fragments, tailored to short-read second generation sequencing, e.g. Illumina, Ion Torrent. To apply the enrichment system to long-read

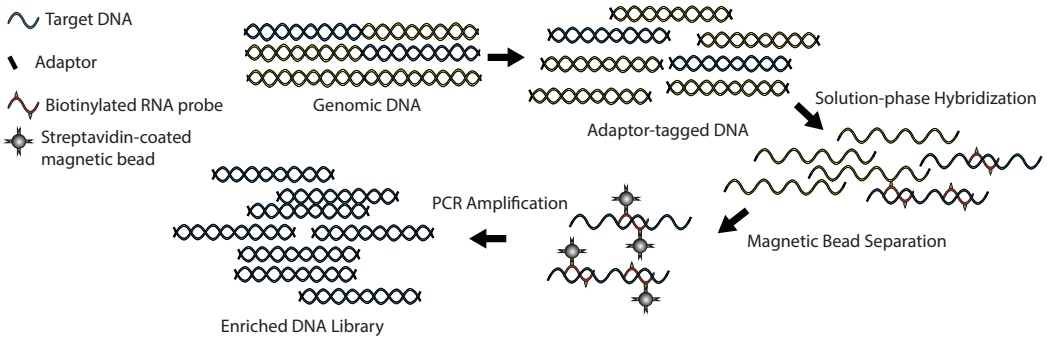


Figure 4.1: Overview of solution-phase hybridization capture. 1) Genomic DNA is sheared and ligated to amplification adaptors. 2) the adaptor-ligated DNA is incubated in solution phase with RNA probes that are complementary to the region of interest and are also biotinylated. 3) Streptavidin-coated magnetic beads are bound to the biotin molecules on the RNA molecules which are in turn hybridized to the region of interest. 4) magnets are used to pull down the streptavidin-coated beads, which selectively retains DNA molecules that are in the region of interest. 5) After washing away non-hybridized DNA, the DNA molecules are amplified by PCR, resulting in enriched DNA pool

sequencing, we adjusted the protocol, altering the shearing conditions to generate DNA fragments with a size distribution centered at 2kb and PCR conditions to allow for amplification of the long DNA strands. The probe design was optimized using Agilent's probe design algorithm and validated experimentally to increase the on-target percentage. Optimizations to the probe design include strategic placement of probes with appropriate, i.e. larger, probe spacing to enrich for larger regions, utilization of stringent probes to decrease non-specific binding, and increased number of probes around regions previously determined to contain SVs.

We applied the modified SureSelectXT protocol on 3-4µg of NA12878 lymphatic cell line gDNA as a control. In addition to the control DNA, we obtained patient-derived pancreatic ductal adenocarcinoma (PDAC) cell lines

and performed the same targeted sequencing on gDNA of these cell lines to examine the performance of the method in real samples. We performed nanopore sequencing on the enriched libraries using Oxford MinION. From an average of 200Mb (100k reads) total sequencing output, we achieved 30 % on-target percentage, yielding an average of >300-fold enrichment in the targeted region (**Figure 4.2, Table 4.1**). To validate the performance of the hybridization capture protocol, the control NA12878 DNA was sheared to 200 bp fragments and selected using the original SureSelect protocol with the same probes, and then sequenced via Illumina short-read sequencing on MiSeq sequencing platform. As shown in Figure 4.2, the alignment coverage over the targeted regions roughly match between the two sequencing platforms.

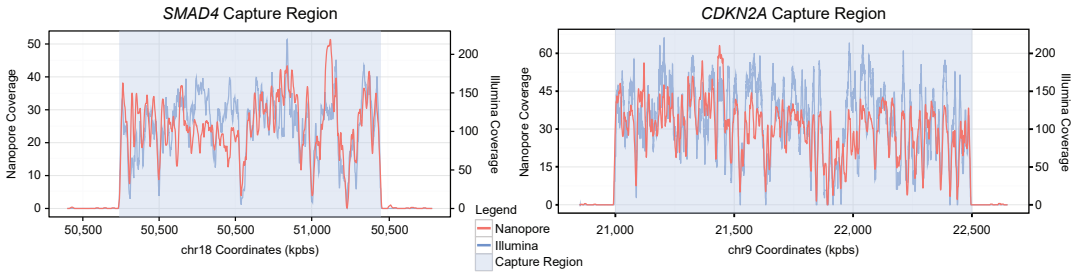


Figure 4.2: Nanopore and Illumina sequencing coverage of the capture regions. Coverages of nanopore and Illumina sequencing using solution-phase hybridization capture system on CDKN2A gene (top) and SMAD4 gene (bottom).

	Total yield (reads)	On-target	On-target percentage	Fold enrichment	Coverage
Illumina NA12878	4.4m	3.7m	85%	641X	113X
Nanopore NA12878	107k	32k	30%	353X	27X
Nanopore PDAC	56k	20k	26%	332X	20X

Table 4.1: Nanopore and Illumina targeted sequencing metrics. Sequencing metrics of Illumina and nanopore sequencing on NA12878 and nanopore sequencing on PDAC cells using solution-phase hybridization capture system.

4.3.2 Detecting nucleotide variations using targeted sequencing

We then detected structural variations (SVs) and single nucleotide variations (SNVs) from the targeted data. We used three approaches in SNV detection to examine whether nanopore sequencing can be effective in detecting SNVs : samtools on Illumina sequencing dataset to serve as the gold-standard approach, nanopore sequencing reads with samtools, and using the raw signal to detect SNVs using nanopolish. Nanopolish corrects errors on the aligned sequences via a hidden Markov Model, wherein the observed output is the k-mer current signal, the states are the true nucleotide sequence, and the conditional probabilities are dependent on the previous state, k-mer current signal, as well as the sequence of the reference alignment (Loman, Quick, and Simpson, 2015). SNV calling on the error-corrected sequences of the control NA12878 yielded 1,017 SNVs, of which 947 were in agreement with the SNVs for the same cell line published via Platinum Genomes Project (Eberle et al., 2016). When compared to the 4,138 SNVs called with raw nanopore data, only 2,485 of which were in agreement with published SNV data, we determined that the majority of the inaccurate SNVs are filtered out through error-correction (**Table** 4.2, example SNVs shown in **Figure** 4.3**a**). SNV analysis of Illumina sequencing data resulted in 1,211 SNVs, of which 1,133 matched the published data. Therefore, with the high depth provided by targeted sequencing, we can achieve the performance of SNV detection in nanopore sequencing that matches NGS. Moreover, using the long reads of nanopore sequencing, we detected de novo heterozygous SNVs and phased

(grouped) the heterozygous SNVs (**Figure 4.3b**).

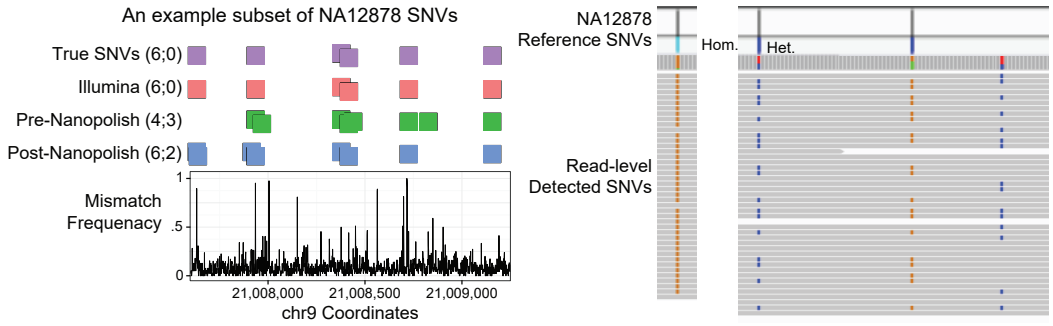


Figure 4.3: SNV analysis using targeted nanopore sequencing. (a) Comparison of detected SNVs and pileup of mismatch in the raw nanopore reads for a 2,000bp window along with mismatch frequencies in the nanopore sequencing data, and **(b)** phased SNV analysis by detecting heterozygous SNVs and phased using the long reads of nanopore sequencing

	Illumina	Pre-polish	Post-polish
Avg. Coverage	113	27	27
Correct	1133	2485	947
Total	1211	4138	1017
Precision	**94%**	60%	**93%**
Sensitivity	**32%**	69%	**26%**

Table 4.2: SNV detection metrics in the targeted regions of interest for three approaches. Calculated precision and sensitivity of Illumina sequencing and the two approaches of nanopore sequencing in detecting single nucleotide variations.

We then detected structural variations using sniffles (Sedlazeck et al., 2018). From the control NA12878 data, 6 SVs were detected in the CDKN2A region and 3 in the SMAD4 region. These SVs were compared to a list of SVs detected via 100x depth whole-genome PacBio long-read sequencing, provided by the Genome In A Bottle consortium, demonstrating that targeted nanopore sequencing captures SVs within the region of interest (**Supplementary Figure**

4.5). We then detected SVs in the PDAC samples, finding two putative SVs, one in each of the two regions (*CDKN2A* and *SMAD4*) (**Figure 4.4**). One SV was present in only a subset of reads, indicating that this SV can be a heterozygous SV. The other SV was a homozygous deletion that spanned a large chunk of the *CDKN2A* (approximately chr9:21,950,000-22,450,00), which was previously discovered by Norris et al (Norris et al., 2015). These SVs on tumor suppressor genes indicate a loss-of-function in these genes, thereby promoting abnormal cellular growth, one of the hallmarks of cancer (Hanahan and Weinberg, 2011).

4.4 Discussion

Targeted sequencing has proven to be a cost-effective method to sequence large number of samples using NGS. We have demonstrated that the same hybridization capture technique used in NGS can be adapted, with modifications, to nanopore sequencing. Using the nanopore hybridization capture sequencing methodology, we have sequenced two tumor suppressor genes to high depths, allowing us to scrutinize genomic anomalies that are present in these genes. When targeting a 2.4 Mbp region (1.5 Mbp for *CDKN2A* gene and 850 kbp for *SMAD4*), which is less than 0.5 % of the genome, 30 % on-target was achieved reliably from nanopore sequencing.

We have shown that this targeted sequencing data can be used to detect SNV and SV detection using nanopolish and sniffles algorithms. With high sequencing depth, the accuracy of SNV calls are dramatically increased and reaches similar performance as Illumina sequencing, the current gold-standard

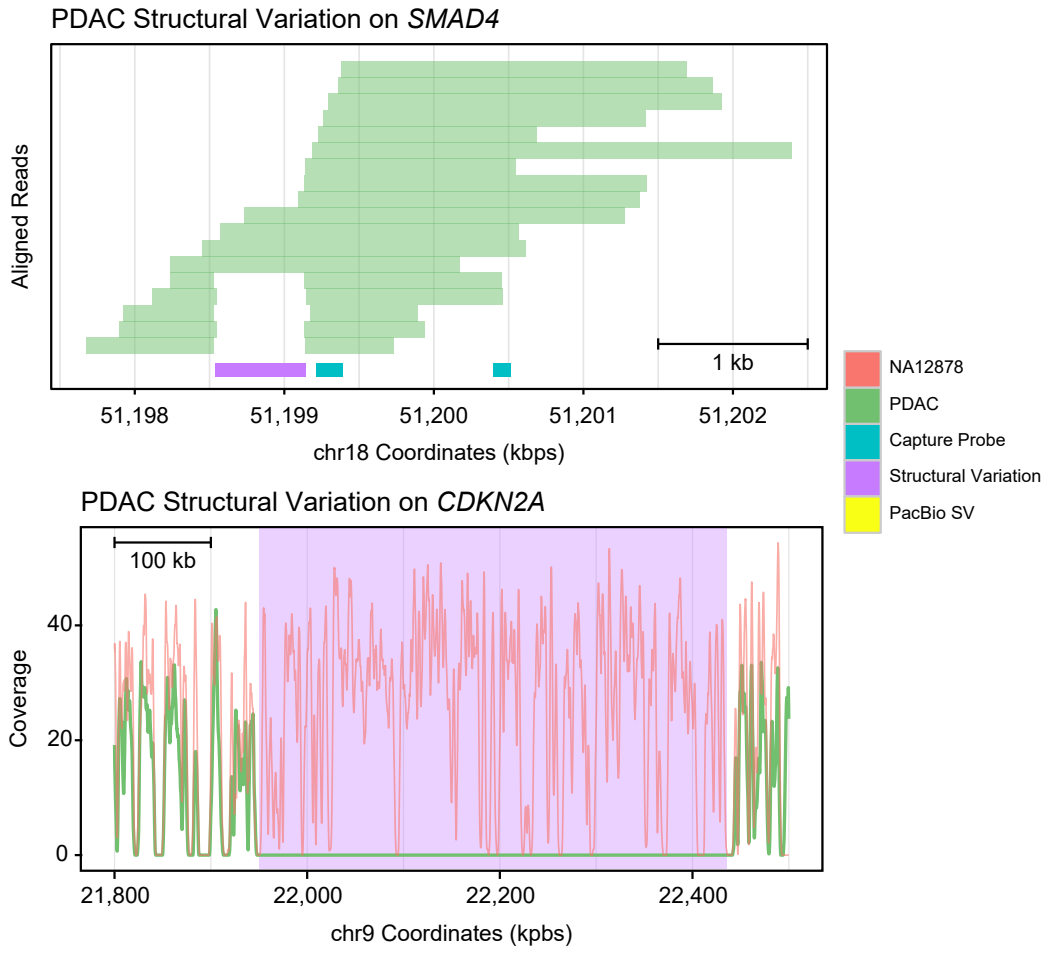


Figure 4.4: Structural Variation detection in targeted nanopore sequencing of Pancreatic Ductal Adenocarcinoma. Structural variations were detected in the targeted nanopore sequencing data of PDAC cell lines

for SNV detection. In addition, the longer reads allows phasing of heterozygous SNVs. As expected, SV detection, which is challenging even with high depth sequencing and intense computational processing from short-read data, is efficient and cost-effective using long-read sequencing coupled with the target enrichment. Applying this technique on patient-derived pancreatic ductal adenocarcinoma cells, we have shown that targeted nanopore sequencing can

be used to detect disease-causing mutations. We have shown that there are SVs in the two targeted tumor suppressor genes, possibly leading to increase in tumor cell growth.

The hybridization capture protocol is currently optimized for 5 kb long reads. While this length is longer than in Illumina sequencing, it is much lower than the length nanopore sequencing is able to handle. Both the phased SNV and SV detection using this technique can be improved with further optimization of the protocol, e.g. enrichment of even longer DNA fragments. In addition, because this protocol enriches for the targeted DNA by PCR amplification, covalent modifications, such as the endogenous CpG methylation, are stripped away. A protocol that removes the need to amplify the DNA will both increase the length of reads as well as allow detection of covalent modifications from nanopore sequencing. In fact, we hav demonstrated the use of the CRISPR-Cas9 system to selectively sequence DNA in the region of interest using nanopore sequencing (Gilpatrick et al., 2020). While the enrichment is not as high as in hybridization capture, this method allows sequencing of long molecules with covalent modifications.

The utility of targeted sequencing has been shown in large scale and clinical studies. The high utility of targeted sequencing is directly translatable to nanopore sequencing : with large scale studies using targeted nanopore sequencing, we will be able to further our understanding of the role of the genome, specifically phased SNVs and larger SVs, and the epigenome in diseases.

4.5 Methods

4.5.1 Agilent Sureselect XT Targeted Enrichment

The enriched DNA library used in nanopore sequencing library preparation was generated using the standard Agilent Sureselect XT protocol with the following modifications. First, the DNA shearing was optimized for fragmentation centering at 5 kb. We used a Bioruptor Pico (Diagenode), shearing 130 µL of 5-50 ng/µL purified genomic DNA in 1.5 mL tubes with 5 cycles of 4 seconds on and 30 seconds off. After end-repair and adaptor ligation, amplification was performed using PCR reagents and conditions optimal for the long fragments. Specifically, we used NEB LongAmp Taq with the following PCR protocol: 30 seconds at 94°C, 8 cycles of 20 seconds at 94°C, 30 seconds at 55°C, and 3 minutes at 65°C, then final extension of 10 minutes at 65°C. The adaptor-ligated 2kb DNA library underwent RNA probe hybridization and capture. Finally, we performed another round of PCR, using NEB LongAmp Taq: 3 minutes at 94°C, 14 cycles of 15 seconds at 94°C, 30 seconds at 60°C, 3 minutes at 65°C, then final extension of 10 minutes at 65°C. As a quality check, we profiled the size distribution and yield on the Bioanalyzer. We used the enriched DNA for Oxford Nanopore Technologies (ONT) DNA sequencing library preparation. Briefly, the DNA fragments were dA-tailed using NEBNext DNA Ultra II reagents and cleaned up with AMPure XP beads. Then, the ONT adaptors were ligated using NEB Blunt/TA ligation master mix. After the final AMPure XP cleanup, the prepared ONT sequencing library was loaded and sequenced per ONT's protocol.

4.5.2 Data preprocessing

Illumina sequencing data was aligned to hg38 human reference genome using bowtie2 with default parameters. Nanopore sequencing data was aligned to the same genome using bwa "mem" module using the pre-tuned options for aligning Oxford Nanopore reads (-x ont2d). In both datasets, samtools was used to sort and convert the resulting data into bam files.

4.5.3 Variant detection using targeted nanopore data

Nanpolish was used to perform error correction using the alignment and the raw sequence. The same set of inputs, along with the output from nanopolish, were used to either call SNVs or build consensus sequence using nanopolish. On Illumina sequencing data and on nanopore sequencing data without error-correction, we used samtools mpileup and bcftools to obtain the putative SNVs. Structural variations were detected using a structural variance caller sniffles using the bwa-mem output bam file, resulting in a variant call format (vcf) file containing loci of putative SVs.

4.6 Acknowledgments

Competing Interests: WT has two patents (8,748,091 and 8,394,584) licensed to Oxford Nanopore. IL has received travel funds to speak at meetings from ONT. The study was partially funded by the supplier of Sureselect system, Agilent Technologies

4.7 Supplementary Material

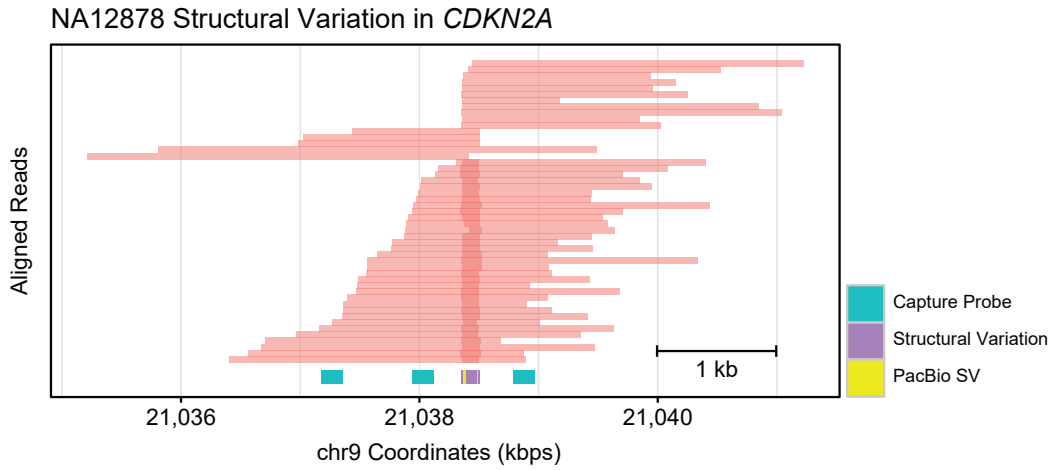


Figure 4.5: Structural Variation detection in targeted nanopore sequencing of NA12878. Structural variations were detected in the targeted nanopore sequencing data of NA12878 and compared with the annotation from the Genome In A Bottle consortium

References

Salzberg, Steven L (2019). "Next-generation genome annotation: we still struggle to get it right". en. In: *Genome Biol.* 20.1, p. 92.

Kozarewa, Iwanka, Javier Armisen, Andrew F Gardner, Barton E Slatko, and C L Hendrickson (2015). "Overview of Target Enrichment Strategies". en. In: *Curr. Protoc. Mol. Biol.* 112, pp. 7.21.1–7.21.23.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts: Laboratory, Data Analysis &Coordinating Center (LDACC): NIH program management: Biospecimen collection: Pathology: eQTL manuscript working group: Alexis Battle, Christopher D Brown, Barbara E Engelhardt, and Stephen B Montgomery (2017). "Genetic effects on gene expression across human tissues". en. In: *Nature* 550.7675, pp. 204–213.

Van Hout, Cristopher V, Ioanna Tachmazidou, Joshua D Backman, Joshua X Hoffman, Bin Ye, Ashutosh K Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Daren Liu, Nilanjana Banerjee, Alexander H Li, O'dushlaine Colm, Anthony Marcketta, Jeffrey Staples, Claudia Schurmann, Alicia Hawes, Evan Maxwell, Leland Barnard, Alexander Lopez, John Penn, Lukas Habegger, Andrew L Blumenfeld, Ashish Yadav, Kavita Praveen, Marcus Jones, William J Salerno, Wendy K Chung, Ida Surakka, Cristen J Willer, Kristian Hveem, Joseph B Leader, David J Carey, David H Ledbetter, Geisinger-Regeneron DiscovEHR Collaboration, Lon Cardon, George

D Yancopoulos, Aris Economides, Giovanni Coppola, Alan R Shuldiner, Suganthi Balasubramanian, Michael Cantor, Matthew R Nelson, John Whittaker, Jeffrey G Reid, Jonathan Marchini, John D Overton, Robert A Scott, Gonçalo Abecasis, Laura Yerges-Armstrong, Aris Baras, and on behalf of the Regeneron Genetics Center (2019). "Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank". en.

Nikiforova, Marina N, Abigail I Wald, Somak Roy, Mary Beth Durso, and Yuri E Nikiforov (2013). "Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer". en. In: *J. Clin. Endocrinol. Metab.* 98.11, E1852–60.

Muller, Kristen E, Jonathan D Marotti, Francine B de Abreu, Jason D Peterson, Todd W Miller, Mary D Chamberlin, Gregory J Tsongalis, and Laura J Tafe (2016). "Targeted next-generation sequencing detects a high frequency of potentially actionable mutations in metastatic breast cancers". en. In: *Exp. Mol. Pathol.* 100.3, pp. 421–425.

Lee, Ji Yun, Kyunghee Park, Sung Hee Lim, Hae Su Kim, Kwai Han Yoo, Ki Sun Jung, Haa-Na Song, Mineui Hong, In-Gu Do, Taejin Ahn, Se Kyung Lee, Soo Youn Bae, Seok Won Kim, Jeong Eon Lee, Seok Jin Nam, Duk-Hwan Kim, Hae Hyun Jung, Ji-Yeon Kim, Jin Seok Ahn, Young-Hyuck Im, and Yeon Hee Park (2015). "Mutational profiling of brain metastasis from breast cancer: matched pair analysis of targeted sequencing between brain metastasis and primary breast cancer". en. In: *Oncotarget* 6.41, pp. 43731–43742.

Loman, Nicholas James, Joshua Quick, and Jared T Simpson (2015). "A complete bacterial genome assembled de novo using only nanopore sequencing data". en.

Eberle, Michael A, Epameinondas Fritzilas, Peter Krusche, Morten Kallberg, Benjamin L Moore, Mitchell A Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J Humphray, Aaron L Halpern, Semyon Kruglyak, Elliott H Margulies, Gil McVean, and David R Bentley (2016). "A reference dataset of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". en.

Sedlazeck, Fritz J, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz (2018). "Accurate detection of complex structural variations using single-molecule sequencing". en. In: *Nat. Methods* 15.6, pp. 461–468.

Norris, Alexis L, Hirohiko Kamiyama, Alvin Makohon-Moore, Aparna Pallavajjala, Laura A Morsberger, Kurt Lee, Denise Batista, Christine A Iacobuzio-Donahue, Ming-Tseh Lin, Alison P Klein, Ralph H Hruban, Sarah J Wheelan, and James R Eshleman (2015). "Transflip mutations produce deletions in pancreatic cancer". en. In: *Genes Chromosomes Cancer*.

Hanahan, Douglas and Robert A Weinberg (2011). "Hallmarks of cancer: the next generation". en. In: *Cell* 144.5, pp. 646–674.

Gilpatrick, Timothy, Isac Lee, James E Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Bradley Downs, Saraswati Sukumar, Fritz J Sedlazeck, and Winston Timp (2020). "Targeted nanopore sequencing with Cas9-guided adapter ligation". en. In: *Nat. Biotechnol.*

Chapter 5

Discussion and Conclusion

I have described my work in using nanopore sequencing to explore the human epigenome. Using exogenous labeling of nuclei with GpC methylation, I simultaneously measured CpG methylation and chromatin accessibility, recapitulating comparable NGS methods and examining parts of the genome that are difficult or impossible to study with conventional NGS methods. With more complete assemblies of the human genome, using the nanoNOMe data directly or other long-read data, we will be able to more accurately examine epigenetic features of repetitive regions and large-scale genomic rearrangements. In fact, Long-read technologies have begun to do so, showing patterns of CpG methylation in centromeric regions of chromosomes and demonstrating the potential significance of the epigenome in these regions (Miga et al., 2019). The ability to assemble the genomes of individual samples using nanoNOMe, resolving genomic variations and their epigenome, will allow more comprehensive understanding of the differences between individuals. We can also use nanopore sequencing to sequence long RNA products of the corresponding cells and integrate with nanoNOMe, directly examining the

relationships of the genome, epigenome, and the transcriptome (Workman et al., 2019).

With rapid advancements of nanopore sequencing technology, additional modifications of DNA beyond GpC methylation could be added to this method (McIntyre et al., 2019). Through the incorporation of additional methyltransferases, (e.g. EcoGII which methylates adenine to N6-methyladenine), it is possible to introduce additional labels in yet another context of the epigenome (Shipony et al., 2020). Such a technique could also provide a "multi-color" measurement, allowing further aspects of the epigenome to be interrogated on the same molecule. Others have already leveraged this methyltransferase fused to lamin protein to explore nuclear architecture but are limited to enzymatic cleavage before sequencing, precluding observation of long-range interactions on a single molecule resolution (Wu, Olson, and Yao, 2016). With further training and development, it may be possible to leverage combinatorial exogenous labeling with nanopore sequencing to ascertain multiple features about chromatin architecture to gain long-range, phased information.

The heterogeneous nature of the epigenome suggests that higher depth would be a key advantage at specific loci. I have shown that targeted sequencing achieves higher depth of sequencing in specific parts of the genome and can be a cost-effective method to study large cohorts of samples. The impact of targeted sequencing would be even greater when coupled with the ability of nanopore sequencing to examine the epigenome. In fact, we have recently developed an improved targeted sequencing method that retains

the modifications on the DNA molecules while achieving >1000X deep sequencing with >20kb long reads (Gilpatrick et al., 2020). With optimizations, nanoNOMe can be streamlined with nCATs, giving us the ability to probe the epigenome of specific regions of the genome with high depth and longer reads. With higher depth, we can more accurately quantify protein binding and promoter epigenetic states and better examine the heterogeneous nature of the epigenome, e.g. differences in protein binding, nucleosome positioning and DNA methylation within a population of cells. The longer reads will allow us to observe cis interactions of protein binding and epigenetic states in more distant loci and understand the association of the epigenetic features in longer distances.

These added capabilities from coupling nanoNOMe - and further extensions of exogenous labeling epigenomic assays - with nCATs will be especially important in observing the epigenome of primary tissue samples. A tissue contains multiple cell types, each of which often have distinct epigenomic profiles (Roadmap Epigenomics Consortium et al., 2015). By resolving multiple epigenetic layers, we can measure the epigenetic heterogeneity in tissue samples and observe the combinatorial epigenetic states in disease-relevant regions on individual DNA strands. This will lead to more comprehensive characterizations of the epigenetic changes during disease progressions and further our knowledge of the mechanisms of diseases.

References

Miga, Karen H, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, Valerie A Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G Bouffard, Alexander M Chang, Nancy F Hansen, Françoise Thibaud-Nissen, Anthony D Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y Dennis, Daniela C Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa Ana Risques, Tina A Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C Mullikin, Pavel A Pevzner, Jennifer L Gerton, Beth A Sullivan, Evan E Eichler, and Adam M Phillippy (2019). "Telomere-to-telomere assembly of a complete human X chromosome". en.

Workman, Rachael E, Alison D Tang, Paul S Tang, Miten Jain, John R Tyson, Roham Razaghi, Philip C Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, Norah Sadowski, Nadine Holmes, Jaqueline Goes de Jesus, Karen L Jones, Cameron M Soulette, Terrance P Snutch, Nicholas Loman, Benedict Paten, Matthew Loose, Jared T Simpson, Hugh E Olsen, Angela N Brooks, Mark Akeson, and Winston Timp (2019). "Nanopore native RNA sequencing of a human poly(A) transcriptome". en. In: *Nat. Methods* 16.12, pp. 1297–1305.

McIntyre, Alexa B R, Noah Alexander, Kirill Grigorev, Daniela Bezdan, Heike Sichtig, Charles Y Chiu, and Christopher E Mason (2019). "Single-molecule sequencing detection of N6-methyladenine in microbial reference materials". en. In: *Nat. Commun.* 10.1, p. 579.

Shipony, Zohar, Georgi K Marinov, Matthew P Swaffer, Nicholas A Sinnott-Armstrong, Jan M Skotheim, Anshul Kundaje, and William J Greenleaf

(2020). “Long-range single-molecule mapping of chromatin accessibility in eukaryotes”. en. In: *Nat. Methods* 17.3, pp. 319–327.

Wu, Feinan, Brennan G Olson, and Jie Yao (2016). “DamID-seq: Genome-wide Mapping of Protein-DNA Interactions by High Throughput Sequencing of Adenine-methylated DNA Fragments”. en. In: *J. Vis. Exp.* 107, e53620.

Gilpatrick, Timothy, Isac Lee, James E Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Bradley Downs, Saraswati Sukumar, Fritz J Sedlazeck, and Winston Timp (2020). “Targeted nanopore sequencing with Cas9-guided adapter ligation”. en. In: *Nat. Biotechnol.*

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, Ginell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis (2015). “Integrative analysis of 111 reference human epigenomes”. en. In: *Nature* 518.7539, pp. 317–330.

Isac Lee
Ph.D. Candidate, Department of Biomedical Engineering
Johns Hopkins School of Medicine
Email: ilee29@jhmi.edu
Github : isaclee

Education

2020	Ph.D. Biomedical Engineering Expected Thesis: *Nanopore Sequencing for Investigation of the Human Epigenome* Mentor: Winston Timp, Ph.D.	Johns Hopkins School of Medicine, Baltimore, MD
2014	B.S. Biomedical Engineering *Mgana Cum Laude*	The University of Texas at Austin, Austin, TX

Professional Experience

2014 –	Graduate Researcher Department of Biomedical Engineering Johns Hopkins School of Medicine, Baltimore, MD *Advisor, Winston Timp*	
2019 – 2019	Graduate Summer Intern Translational Bioinformatics Bristol-Myers Squibb, Lawrenceville, NJ	

Teaching

2017	Department of Biomedical Engineering Introduction to Computing	Johns Hopkins University, Baltimore, MD
2016	Department of Biomedical Engineering Systems Bioengineering III	Johns Hopkins University, Baltimore, MD

Scholarships and Fellowships

2016 – 2017 Nanotechnology for Cancer Research Pre-doctoral Fellowship

2015	Thomas J. Kelly, MD, PhD and Mary L. Kelly Young Scholar Fund

Personal Skills

Bioinformatics	Bash (shell), Python, R, Cloud computing , Amazon web services
Laboratory	Cellular biology, Moleculare biology, epigenetics, DNA sequencing
Languages	English, Korean

Publications

Original Research

[1] T. Gilpatrick, **I. Lee**, J. E. Graham, E. Raimondeau, R. Bowen, A. Heron, B. Downs, S. Sukumar, F. J. Sedlazeck, and W. Timp, "Targeted nanopore sequencing with cas9-guided adapter ligation," *Nat. Biotechnol.*, 2020.

[2] B. A. Avin, Y. Wang, T. Gilpatrick, R. E. Workman, **I. Lee**, W. Timp, C. B. Umbricht, and M. A. Zeiger, "Characterization of human telomerase reverse transcriptase promoter methylation and transcription factor binding in differentiated thyroid cancer cell lines," *Genes Chromosomes Cancer*, vol. 58, no. 8, pp. 530–540, 2019.

[3] **I. Lee**, B. A. Rasoul, A. S. Holub, A. Lejeune, R. A. Enke, and W. Timp, "Whole genome DNA methylation sequencing of the chicken retina, cornea and brain," *Sci Data*, vol. 4, p. 170 148, 2017.

Non-peer Reviewed

[4] S. Aganezov, S. Goodwin, R. Sherman, F. J. Sedlazeck, G. Arun, S. Bhatia, **I. Lee**, M. Kirsche, R. Wappel, M. Kramer, K. Kostroff, D. L. Spector, W. Timp, W. Richard McCombie, and M. C. Schatz, "Comprehensive analysis of structural variants in breast cancer genomes using single molecule sequencing," Nov. 2019, [Online]. Available: https://www.biorxiv.org/content/10.1101/847855v1.

[5] **I. Lee**, R. Razaghi, T. Gilpatrick, N. Sadowski, F. Sedlazeck, and W. Timp, "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing," Dec. 2018, [Online]. Available: https://www.biorxiv.org/content/10.1101/504993v2.

[6] **I. Lee**, R. Workman, J. Wang, and W. Timp, "Use of agilent SureSelect to perform targeted long-read nanopore sequencing," 2017, [Online]. Available: https://www.agilent.com/cs/library/applications/5991-8056EN-2%20Sure%20Select%20App%20Note.pdf.

Conference Talks

[T1] **I. Lee**, R. Razaghi, T. Gilpatrick, N. Sadowski, and W. Timp, "Simulatneous methylation and chromatin accessibility profiling on breast cancer cells," Oxford Nanopore Technologies London Calling, London, UK, May 25, 2018.

[T2] **I. Lee**, R. Razaghi, T. Gilpatrick, N. Sadowski, and W. Timp, "Detecting methylation and chromatin accessibility on long dna sequences," Nuclear Structure and Function, Cold Spring Harbor, NY, May 2, 2018.

[T3] **I. Lee**, R. Razaghi, T. Gilpatrick, N. Sadowski, and W. Timp, "Epigenetic exploration using nanopore sequencing," International Conference on Epigenetics and Bioengineering, Miami, FL, Dec. 14, 2017.

[T4] **I. Lee**, R. Workman, J. Z. Wang, and W. Timp, "Structural variation detection on human dna using targeted sequencing," Oxford Nanopore Technologies Community Meeting, New York, NY, Dec. 2, 2016.

Posters

[P1] **I. Lee**, R. Razaghi, T. Gilpatrick, M. Molnar, N. Sadowski, F. Sedlazeck, K. D. Hansen, J. T. Simpson, and W. Timp, "Methylation and chromatin accessiblility analysis using long-read sequencing," Genome Informatics, Cold Spring Harbor, NY, Nov. 8, 2019.

[P2] **I. Lee**, S. Kumar, N. Majewska, K. McFarland, M. Weinguny, M. Betenbaugh, and W. Timp, "Measuring genomic and epigenomic stability in cho cells: Genome-wide and targeted approaches," Advanced Mammalian Biomanufacturing Innovation Center Biannual Meeting, Boston, MA, Jun. 6, 2019.

[P3] **I. Lee**, R. Razaghi, T. Gilpatrick, N. Sadowski, F. Sedlazeck, and W. Timp, "Nanonome: Profiling methylation and chromatin accessibility simulataneously on individual long dna molecules," Chromatin and Chromosomes Workshop, Baltimore, MD, Dec. 17, 2018.

[P4] **I. Lee**, S. Kumar, N. Majewska, K. McFarland, M. Betenbaugh, and W. Timp, "Understanding and manipulating the egigenome to maximize cho cell productivity and determining genome stability in cho cell lineages," Advanced Mammalian Biomanufacturing Innovation Center Biannual Meeting, St. Louis, MO, Dec. 11, 2018.

[P5] **I. Lee**, R. Workman, J. Z. Wang, and W. Timp, "Targeted nanopore sequencing for variant detection," Advances in Genome Biology and Technology General Meeting, Hollywood, FL, Feb. 15, 2017.

[P6] **I. Lee**, A. Lejeune, and W. Timp, "Unique molecular indexes to remove pcr bias in bisulfite sequencing," The American Society of Human Genetics Annual Meeting, Baltimore, MD, Oct. 8, 2015.

Service and Leadership

2016 – 2017 *Public Relations Representative*, Johns Hopkins University Korean Graduate Student Association

2017 – 2018 *President*, Johns Hopkins University Korean Graduate Student Association